IMPROVED PATIENT IDENTIFICATION AND FEATURE EXTRACTION THROUGH FREE TEXT QUERY AND PROCESSING FOR CLINICAL RESEARCH

by

Douglas Fletcher Redd

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

May 2016

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Douglas Fletcher Redd**

has been approved by the following supervisory committee members:

Bruce E. Bray	, Chair	**16 Feb., 2016** Date Approved
Qing Zeng-Treitler	, Member	**25 Feb., 2016** Date Approved
Gang Luo	, Member	**16 Feb., 2016** Date Approved
Denise E. Beaudoin	, Member	**16 Feb., 2016** Date Approved
Cynthia A. Brandt	, Member	**26 Feb., 2016** Date Approved

and by **Wendy W. Chapman** , Chair/Dean of

the Department/College/School of **Biomedical Informatics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Electronic Health Records (EHRs) provide a wealth of information for secondary uses. Methods are developed to improve usefulness of free text query and text processing and demonstrate advantages to using these methods for clinical research, specifically cohort identification and enhancement.

Cohort identification is a critical early step in clinical research. Problems may arise when too few patients are identified, or the cohort consists of a nonrepresentative sample. Methods of improving query formation through query expansion are described. Inclusion of free text search in addition to structured data search is investigated to determine the incremental improvement of adding unstructured text search over structured data search alone. Query expansion using topic- and synonym-based expansion improved information retrieval performance. An ensemble method was not successful. The addition of free text search compared to structured data search alone demonstrated increased cohort size in all cases, with dramatic increases in some. Representation of patients in subpopulations that may have been underrepresented otherwise is also shown. We demonstrate clinical impact by showing that a serious clinical condition, scleroderma renal crisis, can be predicted by adding free text search.

A novel information extraction algorithm is developed and evaluated (Regular Expression Discovery for Extraction, or REDEx) for cohort enrichment. The REDEx algorithm is demonstrated to accurately extract information from free text clinical

narratives. Temporal expressions as well as bodyweight-related measures are extracted. Additional patients and additional measurement occurrences are identified using these extracted values that were not identifiable through structured data alone. The REDEx algorithm transfers the burden of machine learning training from annotators to domain experts.

We developed automated query expansion methods that greatly improve performance of keyword-based information retrieval. We also developed NLP methods for unstructured data and demonstrate that cohort size can be greatly increased, a more complete population can be identified, and important clinical conditions can be detected that are often missed otherwise. We found a much more complete representation of patients can be obtained. We also developed a novel machine learning algorithm for information extraction, REDEx, that efficiently extracts clinical values from unstructured clinical text, adding additional information and observations over what is available in structured text alone.

TABLE OF CONTENTS

LIST OF FIGURES

Figures

LIST OF TABLES

Tables

CHAPTER 1

INTRODUCTION

1.1 Background

Electronic Health Records (EHRs) provide a wealth of information for secondary uses. EHRs are designed for the primary use of immediate patient care [1], requiring specialized methods to be developed for secondary uses. Some secondary uses include cohort identification for prospective and retrospective clinical research studies and active monitoring for clinical surveillance systems. EHRs generally consist of two categories of data, structured and unstructured. Structured data take the form of distinct values for predefined data points, such as diagnosis codes, vital signs, laboratory results, and demographics. Unstructured data take the form of free text clinical narratives, such as progress notes, discharge instructions, radiology reports, and history and physical reports. Unstructured data are written in narrative form by clinicians, generally with some semi-structured sections interspersed throughout the narrative.

Research and administrative data use from EHRs has primarily been in the form of structured data use. Structured data are relatively accessible and straightforward to interpret. Conversely, unstructured data have been more difficult to access and are much more ambiguous [2]. However, unstructured data are rich in information that is incompletely represented or entirely missing in structured data [3-5]. Searching and interpretation of narrative text has been an active area of research since the earliest days

of computers. Early research in machine translation was overly optimistic, with predictions of complete translation systems by the end of the 1950s [6]. Although these early efforts fell short, much progress has been made since that time. Relatively accurate contemporary systems are available for well-formed language. However, clinical narrative still remains a challenge. Clinical narrative is often not well formed and has many nongrammatical structures. This has given rise to a distinct area of research for the automated processing of clinical narratives [7-9].

The focus of this dissertation is on the use of EHR data in clinical research. In particular, the tasks of cohort identification and enrichment are researched and corresponding methods are developed to demonstrate advantages of free text query and processing (Figure 1.1).

1.2 Cohort Identification

A critical early step in clinical research is the identification of patients who meet inclusion and exclusion criteria for the study. Many problems can arise in this stage, some of which are identifying too few patients for sufficient statistical power, and identifying only a portion of the qualifying population, resulting in a nonrepresentative sample. Traditionally patient sampling has been performed using structured data; however, this has been shown to be insufficient in many cases. Littman et al. [10] found that 32.8% of veterans visiting Department of Veterans Affairs (VA) medical centers in the northwest U.S. did not have height or weight vital signs recorded in their structured data. Kern et al. [11] found that less than 45% of veterans nationwide with diabetes and chronic kidney disease (CKD) had diagnostic codes for CKD. Li et al. demonstrated that a combination of structured and unstructured data sources was useful in identifying

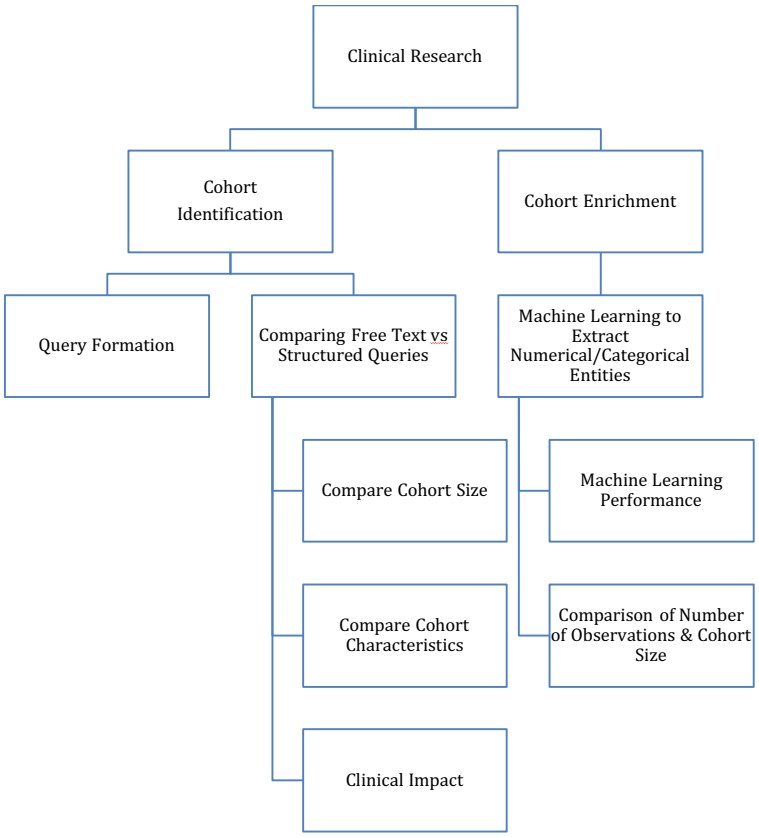

Figure 1.1 Clinical research context

patients with ischemic stroke and coronary artery disease. A more recent study by Abhyankar et al. [12] demonstrated that over half of the ICU patients receiving dialysis in their respective medical systems were only identifiable from their unstructured data. Additionally, some types of data including family history, past medical history, past smoking and alcohol use, and mental health status, frequently are not accommodated in the structured portion of electronic medical record (EMR) systems, and are only available via unstructured data [13].

1.2.1 Query Formation

Many studies begin with an information retrieval step where clinical notes are searched and retrieved. This may be an end in itself, or the first step in a larger analysis

for the purpose of patient identification, such as construction of a cohort of patients experiencing a disease or set of symptoms, or for surveying use of medical procedures in a medical system, etc. Residual error at this stage will cascade to later stages and magnify the error. Identifying patients with specific characteristics from unstructured data can be approached as an information retrieval problem where queries are formulated from a set of keywords. Keyword search is especially effective for rare conditions, which are normally not found in semistructured templates, slot-value sets, or check-boxed lists, and are less common in differential diagnosis sections of documents. It is important to have a broad enough search to produce a result set that is representative of the population of interest. Searches are based on finding documents containing a set of keywords known to be associated with the subject of interest; however, an incomplete set of keywords may result in the omission of relevant documents. Notes containing the keywords indicate an association between the patient to which the clinical note belongs and the characteristic in question. Information retrieval is an active area of research in the medical informatics community as well a much wider community, for example internet searching [14, 15]. Clinical information retrieval has been recognized as having unsolved problems, and was the subject of the medical records track of the Text REtrieval Conference (TREC) focused on methods of accessing free text fields in clinical notes [16-19].

A challenge for keyword-based searching is the creation of a set of keywords that is inclusive enough to find all relevant documents, but specific enough to not include irrelevant documents. Query expansion is a technique that has been used to address this [20, 21]. It has limitations when applied to clinical notes because the vocabulary used can be very different from that in other domains [22-24]. Also, a single expansion method is

generally used which may not be suitable for all situations [25]. In Chapter 2, three separate individual expansion techniques are used, each one with a vocabulary derived from the biomedical domain. In addition, an ensemble method is investigated which integrates the results from the other three in an attempt to improve on the individual methods.

1.2.2 Comparing Free Text and Structured Queries

An essential element of population studies is the identification of groups of patients meeting investigative criteria. Standard database systems used in electronic medical record (EMR) systems support search of structured data, but have very limited support for searching of clinical notes. Search engines such as StarTracker, EMERSE, ARC, and Voogo have been developed to support the use of structured and unstructured data with some success, but have limited usage and require adaptation to different EMR systems from those on which they were developed [26-31]. Identification of patients sharing specific criteria is an essential task for cohort creation. It can be difficult to identify a large enough group of patients for a clinical study, especially in the case of rare or poorly documented conditions. Large EMR systems help in providing a large volume of patients, but this may not be sufficient where the condition in question is not represented in the structured data.

1.2.2.1 Cohort Size

By adding free text query of unstructured data, cohort size can be increased considerably. An important limitation to free text query alone, however, is that retrieval results can include negated mentions of a condition (e.g., “denies smoking”) or mentions

that apply to someone other than the patient (e.g., “mother suffers from chronic kidney disease”). By adding an additional natural language processing (NLP) step, we can address these irrelevant mentions a condition. Systems have been developed for this purpose [13, 32-34] that are highly effective on narrative text, but are limited when used with semistructured text. Maximizing cohort size using unstructured data is explored in detail in Chapter 3, where we identified specific patient cohort requests from researchers, determined the patient population identifiable through query of EMR structured data, and determined the additional patient population that could be identified through free text query and NLP of clinical notes.

1.2.2.2 Cohort Characteristics

When including structured and unstructured data, it is important to consider the possible differences between the patient populations identifiable using those methods. By inclusion of unstructured data, we are able to increase the size of study patient populations and perform studies that may not be possible otherwise. In a large EMR system such as that of the Veterans Health Administration (VHA) with tens of millions of patients, patterns can emerge as to the completeness and methods of documentation of conditions. Patients may be treated in clinics within the same medical system, referred to external providers, or denied treatment depending on many factors including the condition, insurance coverage, geographic accessibility, and regional policy. These patterns of treatment may be represented differentially in structured and unstructured data. For example, a patient receiving treatment from an external provider for a specific condition will normally not have those procedure codes and diagnostic codes reflected in the in-system billing codes. The differences in populations represented by structured and

unstructured data are explored for acupuncture treatments in Chapter 4. In that chapter, we identify a treatment that is underreported in structured data but identifiable in clinical narrative, develop an NLP method for positive identification of the treatment in clinical narrative, and investigate differences in the size and characteristics of the patient populations identified through structured data and NLP of unstructured data.

1.2.2.3 Clinical Impact

It has been shown that inclusion of unstructured data in the search process can considerably increase the number of patients identifiable for cohort inclusion. This method can be extended to domains for which EHR structured data are not suited. Using unstructured data allows us to identify patient cohorts in domains that are largely inaccessible through structured data [10-13]. A challenge in many clinical studies is the under coding of a diagnosis of interest. It is a common practice to identify patients with a clinical condition from their diagnostic billing code, such as the International Classification of Diseases (ICD) versions 9 and 10. Some conditions are underreported, especially when they are secondary diagnoses, do not have adequately matching billing codes, are rare and not widely recognized, or are syndromic, consisting of many co-occurring symptoms and conditions [35-38]. This can become important in cases where a condition has serious counter-indications. These conditions, however, can be identifiable through analysis of the unstructured clinical notes. Both the condition itself, or its co-occurring symptoms and conditions, can be identified and the condition of interest deduced. This can be a powerful tool in detection of conditions that may be only minimally documented or not documented at all in the structured data [39], in which case inclusion of unstructured data enables studies that are not possible otherwise. To illustrate

this, a specific clinical case of identifying patients with risk of scleroderma renal crisis is presented in Chapter 4, where we discover a clinical condition not reliably identifiable through EHR structured data and develop NLP methods for identifying the condition by integrating EHR structured data with unstructured data in clinical narrative.

1.3 Cohort Enrichment

Although EHR systems have made clinical data much more accessible, at times important data points may be missing, incomplete, or of insufficient quality. Data points such as vital signs that are normally stored as structured data are routinely missing in EHR data stores [40]. Other clinically important values are either not represented in structured data or not consistently entered [39, 41]. Compounding this problem, clinicians often enter values as unstructured free text data even when they have the option of entering them as structured data [38, 42, 43].

1.3.1 Machine Learning to Extract Numerical/Categorical Entities

Information extraction from unstructured data enriches the set of clinical values. This is a distinct process from information retrieval, which is more interested in classification. The goal of information extraction is to identify individual data values and measurements within unstructured text. Information extraction has been a topic of research since the early 1970s, and was the subject of the Message Understanding Conferences in the 1980s and 1990s [44]. A common method of extracting values is the use of regular expressions [45]. Computer programmers in association with a clinical expert normally author regular expressions for clinical values [41]. This can become a tedious and brittle process, as hundreds of regular expressions may be required in order to

capture the variety of ways a value may be represented. Sampling is a challenge because any representation not witnessed by the regular expression author is not incorporated into the learned set of regular expressions. Any additions to the corpus necessitate manual review in order to determine if any new representations have been added.

1.3.1.1 Machine Learning Performance

Using a machine learning approach to automatically create regular expressions is an attractive solution. Various algorithms have been developed for this purpose; however, they are limited in their generalizability in that they are task specific [46] or have non-automated steps [47-50]. They are designed for values with very specific patterns that may occur in many contexts. Clinical values, however, are generally indistinguishable from other values by themselves, consisting of numbers, ratios, and ordinal and categorical values. It is only from the context around the value that its meaning can be determined. To address this need, a novel Regular Expression Discovery for Extraction (REDEx) algorithm was developed, with an initial prototype used in Chapter 3, then extended and improved for the use case of temporal expression extraction in Chapter 6. The Regular Expression Discovery for Classification (REDCl) algorithm [51] was adapted for the purpose of extracting clinical values. Temporal expressions such as times, dates, and times relative to events are important in clinical studies in order to establish temporality of events, which is one of the Bradford Hill criteria for causation [52]. Temporal data are straightforward to obtain from structured data, but an important challenge in free text [53, 54].

1.3.1.2 Comparison of Number of Observations and Cohort Size

This has many challenges and complications as investigated in appendix A, where clinical values are identified in unstructured data that are frequently represented in EHR structured data. The values available in the structured data, although frequent, are still incomplete. We adapted the algorithm for application to vital sign values, analyzed the accuracy of the REDEx implementation, and evaluated the magnitude of enrichment of clinical values.

In the following pages, the use of unstructured clinical data for patient identification and feature identification is investigated. A brief overview of the chapter subjects and findings is presented in Table 1.1. Differing methods including types of information retrieval and information extraction are used separately and in combination depending on the task. From keyword based free text search through development of a novel NLP algorithm for clinical value extraction, we demonstrate the use of unstructured data to increase cohort size in order to include patients who may not otherwise be represented, and the ability to perform studies not possible through structured data alone.

Table 1.1 Chapter subjects and findings

Chapter	Research Question	Clinical Domain	Key Findings	Publication	Contribution
2	Can information retrieval in a clinical corpus be improved through query expansion using an ensemble of three biomedically derived query expansion methods?	PTSD Diabetes	The topic model base query expansion performed the best. The ensemble method did not perform as well. A different ensemble approach may be more effective.	Redd D, Rindflesch T, Nebeker J, Zeng-Treitler Q. Improve Retrieval Performance on Clinical Notes: A Comparison of Four Methods. Proceedings of the 46th Hawaii International Conference on System Sciences (HICSS). 2013 2013:2389–97	Collaborated in study design, performed all experiments, primary author of all sections.
3	Can free text query of clinical notes be used to increase the size of clinical cohorts?	Ginkgo/Warfarin Overweight Uncontrolled Diabetes	Cohort size was increased in all cases, especially in Ginkgo/Warfarin where Ginkgo use was sparsely documented in structured data.	Gundlapalli AV, Redd D, Gibson BS, Carter M, Korhonen C, Nebeker J, et al. Maximizing clinical cohort size using free text queries. Computers in Biology and Medicine. 2015 1 May 2015;60:1-7	Collaborated in study design, performed all experiments, primary author of methods and results sections, contributed to all sections.
4	Are there differences between cohorts identified using structured and unstructured data in medical records?	Acupuncture	Cohorts differed in geographic distribution and medical service usage patterns.	Redd D, Kuang J, Zeng-Treitler Q. Differences in Nationwide Cohorts of Acupuncture Users Identified Using Structured and Free Text Medical Records. AMIA Annu Symp Proc. 2014:1002-9	Collaborated in study design, performed all experiments, primary author of all sections.
5	Can patients at risk of complications from complex conditions and interactions be identified using a combination of structured and unstructured data?	Systemic sclerosis Scleroderma renal crisis	Potentially dangerous use of prednisone in patients with systemic sclerosis can be identified using structured data and NLP of unstructured data.	Redd D, Frech TM, Murtaugh MA, Rhiannon J, Zeng QT. Informatics can identify systemic sclerosis (SSc) patients at risk for scleroderma renal crisis. Comput Biol Med. 2014 Aug;53C:203-5. PubMed PMID: 25168254	Designed study, guided and collaboratively performed all experiments, primary author of methods section, contributed to all sections.
6	Can automatic learning of regular expressions reliably identify temporal expressions in free text clinical notes?	Temporality	The REDEx algorithm can be used to automatically learn regular expressions to identify temporal expressions with high sensitivity and specificity.	Redd D, Shao Y, Yang J, Divita G, Zeng-Treitler Q. Automated Learning of Temporal Expressions. Stud Health Technol Inform. 2015;216:639-42	Collaborated in study design, supervised and performed all experiments, primary author of all sections.

Table 1.1 continued

Chapter	Research Question	Clinical Domain	Key Findings	Publication	Contribution
A	Can bodyweight related values be reliably extracted from clinical notes using automatically learned regular expressions?	Bodyweight	Regular expressions can be automatically learned using the REDEx algorithm to extract bodyweight measures with high accuracy.	Murtaugh MA, Gibson BS, Redd D, Zeng-Treitler Q. Regular expression-based learning to extract bodyweight values from clinical notes. Journal of Biomedical Informatics. 2015 April 2015;54:186-90	Collaborated in study design, developed and performed experiments, primary author of methods section, contributed to all sections.

1.4 References

1. Berg M, Goorman E. The contextual nature of medical information. Int J Med Inform. 1999;56(1):51-60.

2. Eisenberg DM, Davis RB, Ettner SL, Appel S, Wilkey S, Van Rompay M, et al. Trends in alternative medicine use in the United States, 1990-1997: results of a follow-up national survey. JAMA. 1998 Nov 11;280(18):1569-75. PubMed PMID: 9820257. Epub 1998/11/20. eng.

3. Lin J, Jiao T, Biskupiak JE, McAdam-Marx C. Application of electronic medical record data for health outcomes research: a review of recent literature. Expert Rev Pharmacoecon Outcomes Res. 2013 Apr;13(2):191-200. PubMed PMID: 23570430. Epub 2013/04/11.

4. Friedlin J, McDonald CJ. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. AMIA Annu Symp Proc. 2006:269-73. PubMed PMID: 17238345. PMCID: 1839544. Epub 2007/01/24.

5. Friedlin J, Grannis S, Overhage JM. Using natural language processing to improve accuracy of automated notifiable disease reporting. AMIA Annu Symp Proc. 2008:207-11. PubMed PMID: 18999177. PMCID: 2656046. Epub 2008/11/13.

6. IBM Archives: 701 Translator 1954 [cited 2015 25 Apr 2015]. Available from: http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html.

7. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128-44. PubMed PMID: 18660887.

8. Li L, Chase HS, Patel CO, Friedman C, Weng C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. AMIA Annu Symp Proc. 2008:404-8. PubMed PMID: 18999285.

PMCID: 2656007. Epub 2008/11/13.

9. Gundlapalli AV, South BR, Phansalkar S, Kinney AY, Shen S, Delisle S, et al. Application of natural language processing to VA electronic health records to identify phenotypic characteristics for clinical and research purposes. Summit on Translat Bioinforma. 2008;2008:36-40. PubMed PMID: 21347124. PMCID: PMC3041527. Epub 2008/01/01. eng.

10. Littman AJ, Boyko EJ, McDonell MB, Fihn SD. Evaluation of a weight management program for veterans. Prev Chronic Dis. 2012;9:E99. PubMed PMID: 22595323. PMCID: PMC3437789. Epub 2012/05/19. eng.

11. Kern EFO, Maney M, Miller DR, Tseng C-L, Tiwari A, Rajan M, et al. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. Health Serv Res. 2006;41(2):564-80.

12. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc. 2014 Sep-Oct;21(5):801-7. PubMed PMID: 24384230. PMCID: PMC4147606. Epub 2014/01/05. eng.

13. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak. 2006 2006;6:30.

14. Kleinberg JM. Authoritative sources in a hyperlinked environment. J Assoc Comput Mach. 1999;46(5):604-32.

15. Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Comput Networks ISDN. 1998 4//;30(1–7):107-17.

16. Joachims T, editor Text categorization with suport vector machines: learning with many relevant features. ECML 1998: Springer-Verlag.

17. Edinger T, Cohen AM, Bedrick S, Ambert K, Hersh W. Barriers to retrieving patient information from electronic health record data: failure analysis from the TREC medical records track. AMIA Annu Symp Proc. 2012 11/03;2012:180-8. PubMed PMID: PMC3540501.

18. Karimi S, Martinez D, Ghodke S, Cavedon L, Suominen H, Zhang L. Search for medical records: NICTA at TREC 2011 Medical Track. TREC 2011.

19. Voorhees EM, Hersh W. Overview of the TREC 2012 medical records track. Available from: http://trec.nist.gov/pubs/trec21/papers/MED12OVERVIEW.pdf.

20. Baeza-Yates R, Ribeiro-Neto B. Relevance feedback and query expansion. Modern Information Retrieval. 2 ed 1999.

21. Manning CD, Raghavan P, Schütze H. Relevance feedback and query expansion. Introduction to Information Retrieval 2008.

22. Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform. 2002 8/2002;35:222-35.

23. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004 Sep-Oct;11(5):392-402. PubMed PMID: 15187068. PMCID: 516246.

24. Zeng QT, Redd D, Divita G, Jarad S, Brandt C, Nebeker JR. Characterizing clinical text and sublanguage: a case study of the VA clinical notes. J Health Med Informat. 2011;4(2). Epub 5.

25. Martinez D, Otegi A, Soroa A, Agirre E. Improving search over electronic health records using UMLS-based query expansion through random walks. J Biomed Inform. 2014;51:100-6. PubMed PMID: 24768598.

26. Hanauer DA. EMERSE: The electronic medical record search engine. AMIA Annu Symp Proc. 2006;2006:941-. PubMed PMID: PMC1839699.

27. D'Avolio LW, Nguyen TM, Farwell WR, Chen Y, Fitzmeyer F, Harris OM, et al. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). J Am Med Inform Assoc. 2010 Jul-Aug;17(4):375-82. PubMed PMID: PMC2995644.

28. Gundlapalli AV, Redd D, Gibson BS, Carter M, Korhonen C, Nebeker J, et al. Maximizing clinical cohort size using free text queries. Comput Biol Med. 2015 1 May 2015;60:1-7.

29. Zeng QT, Redd D, Rindflesch T, Nebeker J. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. AMIA Annu Symp Proc. 2012;2012:1050-9.

30. Redd D, Rindflesch T, Nebeker J, Zeng-Treitler Q. Improve retrieval performance on clinical notes: a comparison of four methods. HICSS 46. 2013:2389–97.

31. Gregg W, Jirjis J, Lorenzi NM, Giuse D. StarTracker: an integrated, web-based clinical search engine. AMIA Annu Symp Proc. 2003;2003:855-. PubMed PMID: PMC1480116.

32. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J

Biomed Inform. 2001 Oct;34(5):301-10. PubMed PMID: 12123149. Epub 2002/07/19. eng.

33. Chapman WW, Chu D, Dowling JN. ConText: an algorithm for identifying contextual features from clinical text. Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing; Prague, Czech Republic. 1572408: Association for Computational Linguistics; 2007. p. 81-8.

34. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 2010-09-06;17:507-13.

35. Ding EL, Song Y, Manson JE, Pradhan AD, Buring JE, Liu S. Accuracy of administrative coding for type 2 diabetes in children, adolescents, and young adults. Diabetes Care. 2007;30(9):e98. PubMed PMID: 17726188.

36. Jordan K, Porcheret M, Croft P. Quality of morbidity coding in general practice computerized medical records: a systematic review. Fam Pract. 2004 Aug;21(4):396-412. PubMed PMID: 15249528

37. Manuel DG, Rosella LC, Stukel TA. Importance of accurately identifying disease in studies using electronic health records. BMJ. 2010;341:c4226. PubMed PMID: 20724404.

38. Ford E, Nicholson A, Koeling R, Tate AR, Carroll J, Axelrod L, et al. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? BMC Med Res Methodol. 2013;13(105).

39. Persell SD, Dunne AP, Lloyd-Jones DM, Baker DW. Electronic health record-based cardiac risk assessment and identification of unmet preventive needs. Med Care. 2009 Apr;47(4):418-24. PubMed PMID: 19238100.

40. Green BB, Anderson ML, Cook AJ, Catz S, Fishman PA, McClure JB, et al. Using body mass index data in the electronic health record to calculate cardiovascular risk. Am J Prev Med. 2012;42(4):342-7. PubMed PMID: PMC3308122.

41. Garvin JH, DuVall SL, South BR, Bray BE, Bolton D, Heavirland J, et al. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. J Am Med Inform Assoc. 2012 Sep-Oct;19(5):859-66. PubMed PMID: 22437073. PMCID: 3422820.

42. Zheng K, Hanauer DA, Padman R, Johnson MP, Hussain AA, Ye W, et al. Handling anticipated exceptions in clinical care: investigating clinician use of 'exit

strategies' in an electronic health records system. J Am Med Inform Assoc. 2011 Nov-Dec;18(6):883-9. PubMed PMID: 21676941. PMCID: 3197991.

43. Zhou L, Mahoney LM, Shakurova A, Goss F, Chang FY, Bates DW, et al. How many medication orders are entered through free text in EHRs?--a study on hypoglycemic agents. AMIA Annu Symp Proc. 2012;2012:1079-88. PubMed PMID: 23304384. PMCID: PMC3540584. Epub 2013/01/11. eng.

44. Grishman R, Sundheim B. Message understanding conference-6: a brief history. Proceedings of the 16th conference on Computational linguistics - Volume 1; Copenhagen, Denmark. 992709: Association for Computational Linguistics; 1996. p. 466-71.

45. Fukuda K-i, Tsunoda T, Tamura A, Takagi T, editors. Toward information extraction: identifying protein names from biological papers. Pac Symp Biocomput; 1998: Citeseer.

46. Prasse P, Sawade C, Landwehr N, Scheffer T. Learning to identify regular expressions that describe email campaigns. ICML, 2012.

47. Li Y, Krishnamurthy R, Raghavan S, Vaithyanathan S, Jagadish HV. Regular expression learning for information extraction. Proceedings of the Conference on Empirical Methods in Natural Language Processing; Honolulu, Hawaii. 1613719: Association for Computational Linguistics; 2008. p. 21-30.

48. Brauer F, Rieger R, Mocan A, Barczynski WM. Enabling information extraction by inference of regular expressions from sample entities. Proceedings of the 20th ACM International Conference on Information and Knowledge Management; Glasgow, Scotland, UK. 2063763: ACM; 2011. p. 1285-94.

49. Babbar R, Singh N. Clustering based approach to learning regular expressions over large alphabet for noisy unstructured text. Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data; Toronto, ON, Canada. 1871848: ACM; 2010. p. 43-50.

50. Xie Y, Yu F, Achan K, Panigrahy R, Hulten G, Osipkov I. Spamming botnets: signatures and characteristics. Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication; Seattle, WA, USA. 1402979: ACM; 2008. p. 171-82.

51. Bui DD, Zeng-Treitler Q. Learning regular expressions for clinical text classification. J Am Med Inform Assoc. 2014 Sep-Oct;21(5):850-7. PubMed PMID: 24578357. PMCID: PMC4147608.

52. Hill AB. The environment and disease: association or causation? Proc R Soc Med. 1965 May;58(5):295-300. PubMed PMID: 14283879.

53. Zhou L, Hripcsak G. Temporal reasoning with medical data - a review with emphasis on medical natural language processing. J Biomed Inform. 2007;40(2):183-202.

54. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc. 2013;20(5):806-13.

CHAPTER 2

IMPROVE RETRIEVAL PERFORMANCE ON CLINICAL NOTES: A COMPARISON OF FOUR METHODS

2013 46th Hawaii International Conference on System Sciences

Improve Retrieval Performance on Clinical Notes: A Comparison of Four Methods

Doug Redd
VA Salt Lake City Health Care System, UT, USA
Department of Biomedical Informatics, University of Utah, USA
doug.redd@utah.edu

Thomas Rindflesch
National Library of Medicine, Bethesda, MD, USA
trindflesch@mail.nih.gov

Jonathan Nebeker
Geriatric Research, Education, and Clinical Center (GRECC), VA Salt Lake City Health Care System, UT, USA
Dept. of Internal Medicine, University of Utah, USA
jonathan.nebeker@hsc.utah.edu

Qing Zeng-Treitler
VA Salt Lake City Health Care System, UT, USA
Department of Biomedical Informatics, University of Utah, USA
q.t.zeng@utah.edu

Abstract

Query expansion is a commonly used approach to improving search results. Specific expansion methods, however, are expected to have different results. We have developed three different expansion methods using knowledge derived from medical thesaurus, medical literature, and clinical notes. Since the three different sources each have strengths and weaknesses, we hypothesized that combining the three sources will lead to better retrieval performance. Evaluation was performed for the 3 different query expansion techniques and an ensemble method on two sets of clinical notes. 11-point interpolated average precisions, MAP, and P(10) scores were calculated which indicate that topic model based expansion has the best results and the predication method the worst. This finding points to the potential of the topic modeling methods as well as the challenge in integrating different knowledge sources.

1. Introduction

Unstructured text contained in medical and clinical notes includes valuable information that is often not available through associated structured data. Significant effort and resources are dedicated to the creation and storage of these records, the generation of which is accelerating to fill large data stores. Each patient may have hundreds of unstructured text notes. This information is of use to many areas outside of the medical clinic. Researchers may search for specific patient populations while administrators may seek collection of performance measures [16, 17]. Quick and accurate retrieval has become a significant challenge.

Initially the challenge can appear to be solvable with traditional text search, however this is often ineffective. Searching for "Multiple Sclerosis", for example, will miss occurrences where it is commonly abbreviated to "MS". However, adding "MS" to the search will also result in many false positives (MS is used commonly in clinical documents to indicate an adult female, milliseconds, morphine sulfate, etc.). False negatives can also occur when the disease is referred to by another name, such as disseminated sclerosis.

A frequent method for improving performance of a search comes from the field of information retrieval (IR) in the form of query expansion [10]. This method consists of adding additional related query terms to the search. These additional terms are commonly obtained from a thesaurus or from records of previous related searches [5, 31]. A certain algorithm is then used to rank the retrieved documents in order of relevancy. The process of query term addition may be automatic or it may be interactive, where the additional terms are presented to the user, who then chooses which additional terms to include [14]. Many query expansion methods have been in biomedical informatics in literature searching applications [1, 9, 15, 25, 27, 30]. Several studies have used clinical notes to investigate different query expansion methods with mixed results.

This study describes our experiments contrasting three different query expansion techniques with an ensemble method of combining the results of all three techniques into a single result. The three query expansion techniques are based on synonyms, topic

1530-1605/12 $26.00 © 2012 IEEE
DOI 10.1109/HICSS.2013.261

models, and predications. In the synonym technique, additional query terms consist of synonyms identified in a subset of the UMLS vocabularies along with their lexical variants. In the topic model technique, additional query terms consist of related words from a topic model that was created from 100,000 clinical notes. In the predication technique, additional query terms are generated from a predication database that was created from medical literature using SemRep (a natural language processing (NLP) system). In the ensemble technique, additional terms are taken from all 3 of the other techniques, the terms are then ranked, and the most relevant ones are used for query expansion.

2. Background

Most of biomedical IR research has focused on literature rather than clinical notes. Prior research in clinical IR has predominantly been in the areas of query log analysis [23, 24, 34], temporal relationships [2, 6, 8, 19], ontology and terminology based query expansion [20, 29], and bundled query sets [13]. From the results recent studies testing IR techniques on unstructured text, evidence has shown that user queries need to be expanded, however specific expansion techniques such as relevance feedback and concept indexing are not universally effective [4, 7, 18, 21, 22, 26]. Term weighting has been demonstrated to improve document ranking, and applying NLP techniques to filtering has been shown to improve precision.

2.1. Synonym based expansion

A number of studies have investigated the use of synonyms for expansion with mixed results. One such experiment showed a minor increase in recall by applying UMLS synonyms, with improved results being obtained when restricting to the MeSH vocabulary [12]. Another study showed degradation in query performance when using synonyms [15]. Some studies from TREC 2011 did not demonstrate any performance gain when using synonyms. We have restricted our synonym expansion process to a subset of UMLS source vocabularies in an attempt to optimize expansion performance.

2.2. Topic model based expansion

In this context, a topic can be defined as a collection of words or terms that frequently occur together and are related to the same subject. Topic modeling analyzes the patterns of words, terms, or concepts in a corpus to elucidate common topics. In this regard it is related to latent semantic analysis (LSA). However, while LSA attempts to discover the relationships between concepts, topic modeling attempts to discover the hidden thematic structure in a document or corpus. An unsupervised method can be used with topic modeling to obtain topics from a text corpus [32]. This approach was used by Steyvers and Griffiths for analysis of abstracts from the Proceedings of the National Academy of Sciences [11]. They found the extracted topics represented meaningful structure in the abstracts that was consistent with class designations provided by the authors of the articles. We ourselves had performed a couple of experiments using sample-based implementations of Latent Dirichlet Allocation (LDA) [3, 15] to test the use of topic modeling in IR [36].

2.3. Predication based expansion

We employed a predication database in this study that was constructed by applying the SemRep NLP system [28] to MEDLINE citations. SemRep identifies semantic propositions (predications) in biomedical documents. It uses the MetaMap system to assign semantic types to noun phrases, the SPECIALIST lexicon for lexical variant identification, and the Xerox part-of-speech tagger. Evaluation studies of SemRep report it's precision to be approximately 75%. The database of predications we used contains 25 million predications from ten years of MEDLINE (1999-2009).

2.4. Ensemble expansion

In this technique, we merge the highest ranked related terms from the Synonym, Topic Modeling, and Predication techniques to assemble a combined set of expanded terms. To merge the related terms we calculated the overall semantic distance between the query terms and the normalized ranks from the other techniques.

3. Methods

We will describe the data sets used in the evaluation, details of the four query expansion methods, and then our methods of scoring and evaluation.

3.1. Data Sets

Data sets were obtained from the VINCI database that contains structured and unstructured data. VINCI is an initiative to promote analysis of VA data by improving access for researchers while ensuring data privacy and security [33]. VINCI is partnered with the VA Corporate Data Warehouse (CDW) and hosts data available through CDW as well as some other data sources, currently containing over 1.8 billion clinical notes for over 30 million patients. Two data sets were used for evaluation, a PTSD data set and a Diabetes data set. The PTSD data set was taken from patients with two or more ICD9 codes of 309.81, while the Diabetes data set was taken from patients with two or more ICD9 codes of 250.*. Each data set consisted of 300 clinical notes randomly selected from the identified patient sets.

3.2. Synonym based expansion

We identified synonym using the UMLS. We first mapped query terms to UMLS concepts, if possible, using MetaMap. If no concept mappings were found using MetaMap, concepts were assigned using a term-to-concept table derived from MRCONSO 2011AA. To reduce the rate of uninformative concepts, we restricted the data sources to SNM, SNOMEDCT, MSH, and ICD. Once query terms were mapped to concepts, those concepts were used to look up related concepts in the MRREL table. Related concepts were ranked by frequency, with individual weights being determined by relationship type. We assigned "child" relationships a weight of 2, "not related, no mapping", "allowed qualifier", and "can be qualified by" relationships a weight of 0, and all other relationships a weight of 1.

3.3. Topic model based expansion

We used the MALLET program to identify 1,000 topics using implementations of the Latent Dirichlet Allocation (LDA), Pachinko Allocation, and Hierarchical LDA algorithms. Topic models are assembled by assigning the words of each document to one of a number of topics. It is used to automatically group words into meaningful topics. A sample topic is shown in Table 1.

Topics were generated from a corpus of 100,000 clinical documents from the VINCI dataset. The corpus was assembled by determining the 100 most frequent document types in the VINCI dataset, then selecting 1,000 documents from each of those document types. The topics containing the query terms were then identified, and related terms in those topics were selected. The related terms were ranked by relative weight in the topic and overall topic weight.

Table 1. Sample Topic

Word	Probability of word occurrence in the topic
Diabetes	0.56
Insulin	0.37
Glucose	0.21
HbA1C	0.09
Hypertension	0.01

3.4. Predication based expansion

Subjects of predications where the object was one of the query terms were found in a predications database. Those matching subjects were ranked by the frequency of co-occurrence with query term objects. The predication database was created using the SemRep program with MEDLINE abstracts as the source documents. No restrictions were placed on the type of predicate.

3.5. Ensemble expansion

To combine the results of the sources for expansion, the scores of each result set were first normalized to values between 0 and 1 using the equation:

$$\text{if } Score(C_x, C_y, s) > 0$$
$$i_s(C_x, C_y, s) = \frac{\ln\left(Score(C_x, C_y, s)\right) + 1}{\ln\left(MAX(Score(C_x, C_n, s))\right) + 1}$$
$$\text{if } Score(C_x, C_y, s) = 0$$
$$i_s(C_x, C_y, s) = 0$$

where

C_x is the query term
C_y is a related term
C_n is any related term
s is the source of the expansion

The normalized result sets were then combined and the resulting term scores were determined by calculating their semantic distance:

$$Semantic\ Distance = (i_{A \cap B \cap C} \times 1000) + i_{A \cup B \cup C}$$

where

$$i_{A \cap B \cap C} = i_A * i_B * i_C$$

$$i_{A \cup B \cup C} = i_A + i_B + i_C - (i_A * i_B) - (i_A * i_C) - (i_B * i_C) + (i_A * i_B * i_C)$$

are the fuzzy intersection and union of the result sets from the 3 expansion sources.

3.6. Query Expansion Scoring

Two document sets were used in scoring the expansions, one being a set of documents related to PTSD and the other related to diabetes. The two document sets consisted of 300 VINCI TIU documents each selected randomly from patients with ICD9 codes of 309.81 and 250. To facilitate query expansion scoring, TF-IDF scores were calculated for every term in the two document sets. Query expansions were scored for each document by summing the individual TF-IDF scores of the expanded terms in each document.

Table 2. Examples of synonym based expansion, topic model based expansion, predication based expansion, and ensemble expansion

Original query	education, cognitive, behavioral, therapy	diet, exercise
Synonym based expansion	accident, behavior, qualifier, compulsion, education, proneness, general, habits, modifier, cognitive, qualifiers, social, behavioral, sexual, therapy	nutritional, diet, diets, exercise, fasting, vegetarian, diabetic, find, nutrition, dietary, supplement, conditioning, diet
Topic model based expansion	condition, functioning, learn, learning, education, barriers, cognitive, emotional, mental, behavioral, provided, therapy, advance	control, nutrition, counseled, intake, inr, diet, fat, exercise, psa, warfarin, cholesterol, wt
Predication based expansion	control, university, immune, transplantation, implant, available, education, surgery, destination, improved, cognitive, surgical, behavioral, compression, therapy	supplementation, safflower, diet, exercise, regimen, regimens, manipulation, oil, complete, soybean, supplemented, salt, treatments
Ensemble expansion	method, sib, treatment, behaviour, immunotherapy, occupational, self, education, median, cognitive, school, behavioral, therapies, therapy, injurious	exercise, dietary, fat, intake, diet, exercises, exercised, diets, sustained, calorie, ambiguous, nutrition

3.7. Evaluation

Six PTSD and six Diabetes queries were manually evaluated by a clinician against each document in the two data sets. The queries are show in Table 3. A Likert scale was used in ranking each document: 0: irrelevant, 1: possibly relevant, and 2: definitely relevant. The category of "possibly relevant" is needed because at times the context of the document may indicate relevance without making it clear. This may happen in a document where symptoms of a disease are mentioned but not the disease itself, for example.

We applied three query expansion techniques, a fourth technique that combined the expansions of the other three techniques, and a baseline without query expansion for a total of 5 expansion techniques. Six PTSD queries were used for the PTSD document set and six Diabetes queries were used for the Diabetes document set. We used TF-IDF scoring to rank the documents retrieved by each of the five expansion techniques.

We calculated the 11-point interpolated average precision (IAP) for each query. This is a standard method of evaluating ranked results in IR. We ranked each document by TF-IDF, then calculated precision and recall at each TF-IDF ranking in the document set. Interpolated precision was determined at each of the 11 recall points by taking the maximum precision for recalls greater than each recall point. A composite 11-point IAP was calculated across all queries for overall comparison of expansion methods.

We determined the mean average precision (MAP) score by averaging the precisions at the points where each relevant document was retrieved. We determined the precision at 10 (P(10)) score by taking the precision at the point where the tenth document was retrieved. The average P(10) score for each query was then calculated.

Table 3. Query definitions

Query set	Query label	Query terms
PTSD	PTSD query	PTSD
	Suicide ideation query	suicide, ideation, homicide, SI, HI
	Medication query	psych, PTSD, medication, drug
	Change query	better, worse, change
	Symptom query	PTSD, symptoms, intrusive, upsetting, memories, flashbacks, nightmares, intense, distress, reactions, reminders, event
	Therapy query	education, cognitive, behavioral, therapy
Diabetes	Diabetes query	diabetes
	Neuropathy query	wound, neuropathy
	Medication query	diabetes, medication, drug
	Change query	better, worse, change
	Symptom query	diabetes, symptoms, frequent, infections, blurred, vision, cuts, bruises, slow, heal, tingling, numbness, hands, feet, recurring, skin, gum, bladder, infection
	Therapy query	diet, exercise

4. Results

We calculated the 11-point interpolated average precisions for each query and plotted precision-recall curves. We then calculated the composite 11-point IAP averaged across all of the queries. Mean Average Precision (MAP) and Average Precision at 10 (P(10)) scores were then calculated for each expansion method. The plot with the largest area under the curve indicates the best retrieval method because it represents overall higher precision and recall. Although the individual curves are highly varied, the composite IAP indicates the topic model and synonym curves outperform the ensemble and predication methods, although all methods outperformed the baseline. The MAP and average P(10) scores also show the topic model and synonym methods outperforming the other methods.

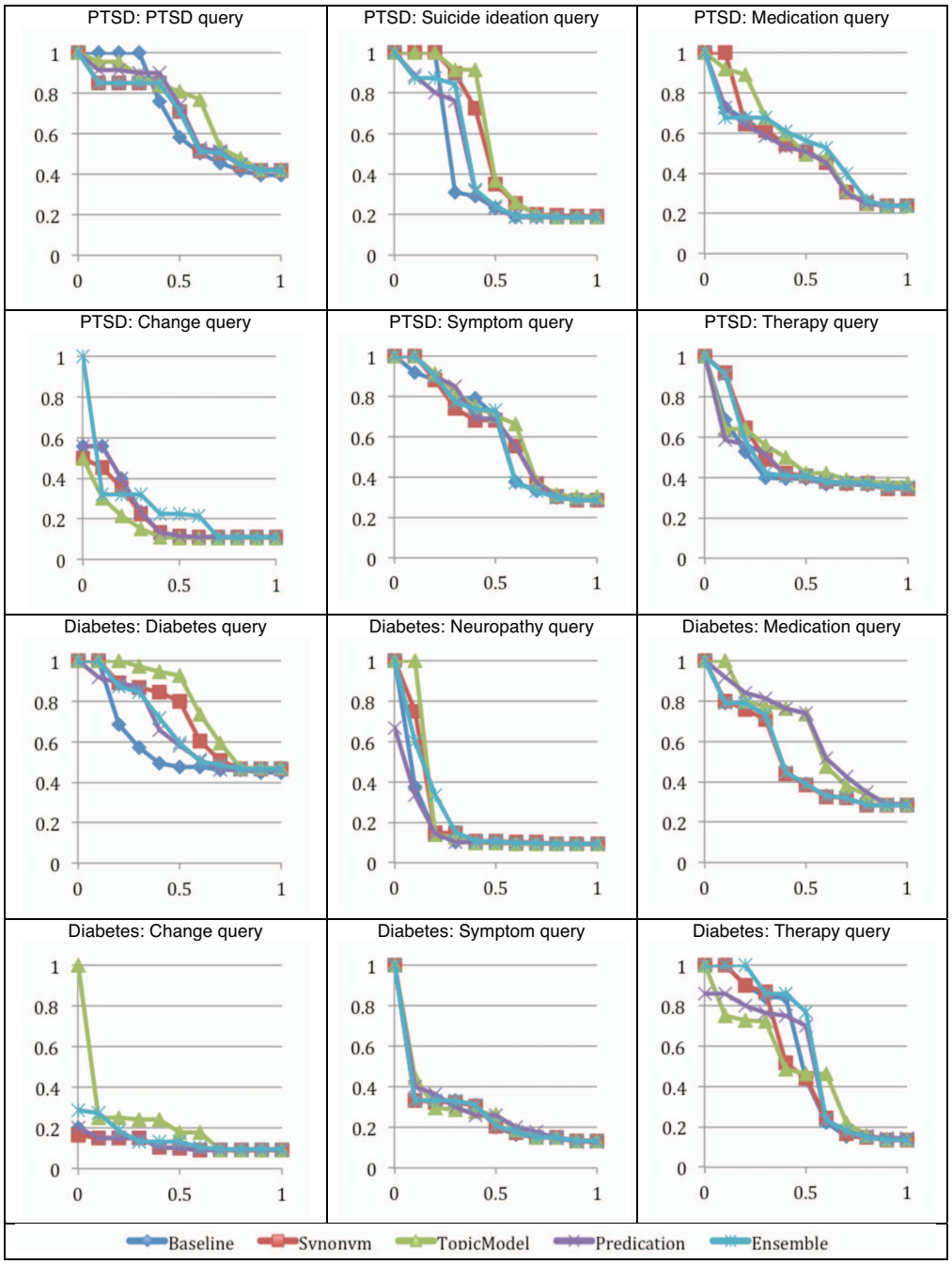

Figure 1. 11-point IAP for PTSD and diabetes queries. X-axis is recall, Y-axis is int. precision

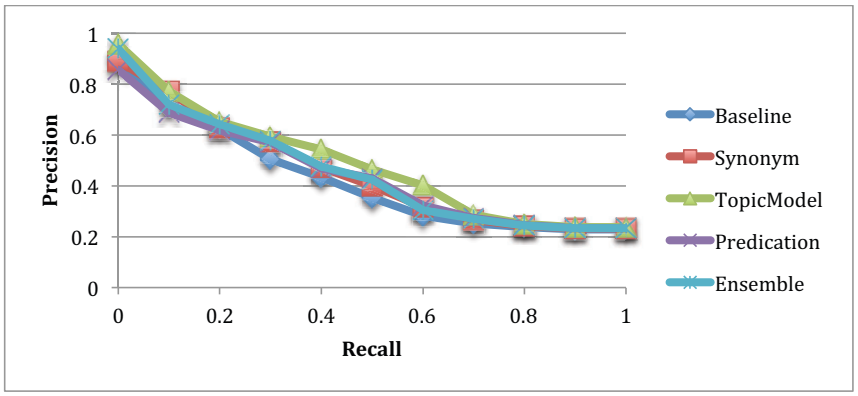

Figure 2. Composite 11-point IAP of all queries

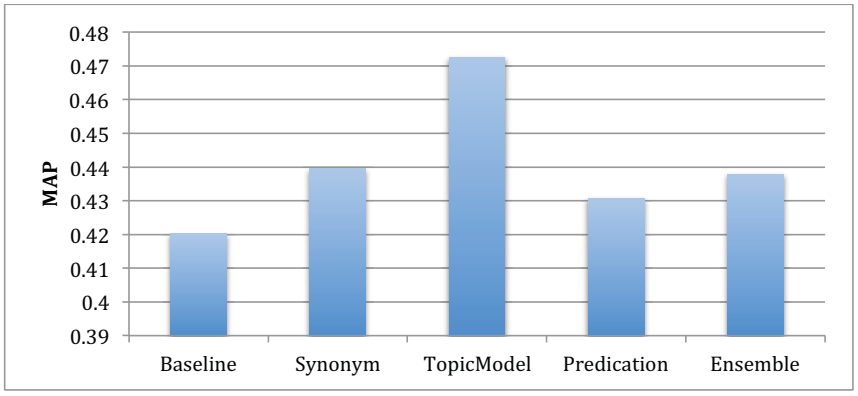

Figure 3. Mean Average Precision (MAP) of expansion techniques

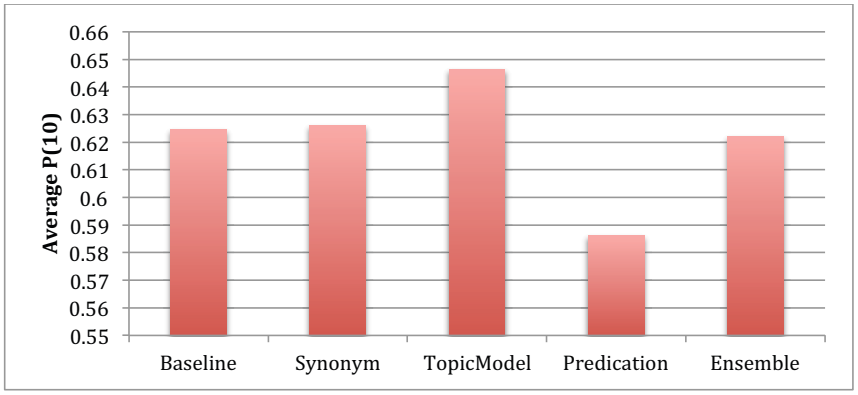

Figure 4. Average P(10) of expansion techniques

5. Discussion

Four methods were used in the expansion of twelve queries over a corpus of 600 documents. One of the methods was an ensemble, which was introduced as a possible way of combining the best of the other methods. The topic model and synonym methods outperformed the baseline. The ensemble and predication methods outperformed the baseline in all but the P(10) score. A danger of query expansion is that it can sacrifice precision in the quest to increase recall, and it is worth noting that these expansions did not have a significantly negative effect on precision overall.

There is large variance in the four methods' relative performance among the difference queries and targeted level of precision and recall. As shown in Figure 1, each method performed well on certain queries and poorly on others. The variance is understandable since some query terms have large numbers of synonyms, while others have large numbers of closely related terms. When there are fewer synonyms, for example, the benefit of synonym expansion will be limited. If a synonym or related term happens to be ambiguous, the false negatives may also reduce the usefulness of the query expansion.

There are two potential solutions. One is to develop interactive methods and avoid automated expansions. Indeed, a user would be the best judge of the appropriateness of the expansion terms. The users of clinical datasets are likely to be willing to spend some time interacting with the query system since the quality of the retrieved datasets is likely to impact the later analysis steps which will take hours, days, even months.

Another solution is to develop more sophisticated expansion algorithms. Our first intuition was that combining multiple knowledge sources would lead to better results. Though this is true for a few queries, the overall performance of our ensemble algorithm was not satisfactory. Ensemble systems have been shown previously to successfully outperform individual methods [35]. One possible cause of the ensemble system underperforming the topic model and synonym systems is that the semantic distance measure we used gives precedence to overlapping terms. In the case of query expansion this may penalize performance.

Using high ranking but non-overlapping terms may give better results in the ensemble system. This might be accomplished by selecting individually highly ranked search terms from the other expansions and avoiding overlapping terms rather than calculating semantic distance. Another improvement may come from enriching the topic model expansion. Using it as a starting point, retaining all of its expansions, then adding only highly ranked, non-overlapping expansion terms from the other methods.

6. Acknowledgement

This work is funded by VA grants CHIR HIR 08-374 and VINCI HIR-08-204.

7. References

[1] Aronson, A.R., and Rindflesch, T.C., "Query Expansion Using the Umls Metathesaurus", Proc AMIA Annu Fall Symp, 1997, pp. 485-489.

[2] Bellika, J.G., Sue, H., Bird, L., Goodchild, A., Hasvold, T., and Hartvigsen, G., "Properties of a Federated Epidemiology Query System", Int J Med Inform, 76(9), 2007, pp. 664-676.

[3] Blei, D.M., Ng, A.Y., and Jordan, M.I., "Latent Dirichlet Allocation", Journal of Machine Learning Research, 3(2003, pp. 993-1022.

[4] Chung, J., and Murphy, S., "Concept-Value Pair Extraction from Semi-Structured Clinical Narrative: A Case Study Using Echocardiogram Reports", AMIA Annu Symp Proc, 2005, pp. 131-135.

[5] Cui, H., Wen, J.-R., Nie, J.-Y., and Ma, W.-Y., "Query Expansion by Mining User Logs", in (Editor, 'ed.'^'eds.'): Book Query Expansion by Mining User Logs, 2003, pp. 829-839.

[6] Deshpande, A.M., Brandt, C., and Nadkarni, P.M., "Temporal Query of Attribute-Value Patient Data: Utilizing the Constraints of Clinical Studies", Int J Med Inform, 70(1), 2003, pp. 59-77.

[7] Dinh, D., and Tamine, L., "Irit at Trec 2011: Evaluation of Query Expansion Techniques for Medical Record Retrieval.
", Text Retrieval Conference,TREC 2011, 2011

[8] Dorda, W., Gall, W., and Duftschmid, G., "Clinical Data Retrieval: 25 Years of Temporal Query Management at the University of Vienna Medical School", Methods Inf Med, 41(2), 2002, pp. 89-97.

[9] Doszkocs, T.E., "Aid, an Associative Interactive Dictionary for Online Searching", Online Review, 2(2), 1978, pp. 163-172.

[10] Frakes, W.B., and Baeza-Yates, R., Modern Information Retrieval, Addison Wesley, 1999.

[11] Griffiths, T.L., and Steyvers, M., Finding Scientific Topics, National Acad Sciences, 2004.

[12] Griffon, N., Chebil, W., Rollin, L., Kerdelhue, G., Thirion, B., Gehanno, J.-F., and Darmoni, S., "Performance Evaluation of Unified Medical Language System®'S Synonyms Expansion to Query Pubmed", BMC Medical Informatics and Decision Making, 12(12), 2012,

[13] Hanauer, D.A., "Emerse: The Electronic Medical Record Search Engine", AMIA Annu Symp Proc, 2006, pp. 941.

[14] Harman, D., "Towards Interactive Query Expansion", in (Editor, 'ed.'^'eds.'): Book Towards Interactive Query Expansion, ACM Press, Grenoble, France, May 1988, pp. 321-331.

[15] Hersh, W., Price, S., and Donohoe, L., "Assessing Thesaurus-Based Query Expansion Using the Umls Metathesaurus", Proc AMIA Symp, 2000, pp. 344-348.

[16] Horvath, M.M., Winfield, S., Evans, S., Slopek, S., Shang, H., and Ferranti, J., "The Deduce Guided Query Tool: Providing Simplified Access to Clinical Data for Research and Quality Improvement", J Biomed Inform, 44(2), 2011, pp. 266-276.

[17] Hu, H., Correll, M., Kvecher, L., Osmond, M., Clark, J., Bekhash, A., Schwab, G., Gao, D., Gao, J., Kubatin, V., Shriver, C.D., Hooke, J.A., Maxwell, L.G., Kovatich, A.J., Sheldon, J.G., Liebman, M.N., and Mural, R.J., "Dw4tr: A Data Warehouse for Translational Research", J Biomed Inform, 44(6), 2011, pp. 1004-1019.

[18] Karimi, S., Martinez, D., Ghodke, S., Zhang, L., Suominen, H., and Cavedon, L., "Search for Medical Records:Nicta at Trec 2011 Medical Track", TREC 2011, 2011

[19] Klimov, D., Shahar, Y., and Taieb-Maimon, M., "Intelligent Interactive Visual Exploration of Temporal Associations among Multiple Time-Oriented Patient Records", Methods Inf Med, 48(3), 2009, pp. 254-262.

[20] Mabotuwana, T., and Warren, J., "An Ontology-Based Approach to Enhance Querying Capabilities of General Practice Medicine for Better Management of Hypertension", Artif Intell Med, 47(2), 2009, pp. 87-103.

[21] Mariam Daoud, D.K., Jun Miao, Jimmy Huang, "York University at Trec 2011: Medical Records
Track", TREC 2011, 2011

[22] Martijn Schuemie, D.T., Edgar Meij, "Dutchhattrick: Semantic Query Modeling, Context,
Section Detection, and Match Score Maximization", TREC 2011, 2011

[23] Murphy, S.N., Morgan, M.M., Barnett, G.O., and Chueh, H.C., "Optimizing Healthcare Research Data Warehouse Design through Past Costar Query Analysis", Proc AMIA Symp, 1999, pp. 892-896.

[24] Natarajan, K., Stein, D., Jain, S., and Elhadad, N., "An Analysis of Clinical Queries in an Electronic Health Record Search Utility", Int J Med Inform, 79(7), 2010, pp. 515-522.

[25] Nenadic, G., Mima, H., Spasic, I., Ananiadou, S., and Tsujii, J., "Terminology-Driven Literature Mining and Knowledge Acquisition in Biomedicine", Int J Med Inform, 67(1-3), 2002, pp. 33-48.

[26] Ozturkmenoglu, O., and Alpkocak, A., "Demir at Trec-Medical 2011: Power of Term Phrases in Medical Text Retrieval", 20th Anniversary of Text Retrieval Conference, 2011

[27] Pollitt, S., "Cansearch: An Expert Systems Approach to Document Retrieval", Information Processing & Management, 23(2), 1987, pp. 119-138.

[28] Rindflesch, T.C., and Fiszman, M., "The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text", J Biomed Inform, 36(6), 2003, pp. 462-477.

[29] Schulz, S., Daumke, P., Fischer, P., and Muller, M., "Evaluation of a Document Search Engine in a Clinical Department System", AMIA Annu Symp Proc, 2008, pp. 647-651.

[30] Srinivasan, P., "Retrieval Feedback in Medline", J Am Med Inform Assoc, 3(2), 1996, pp. 157-167.

[31] Stenmark, D., "Query Expansion on a Corporate Intranet:
Using Lsi to Increase Precision in Explorative Search", in (Editor, 'ed.'^'eds.'): Book Query Expansion on a Corporate Intranet:
Using Lsi to Increase Precision in Explorative Search, 2005, pp. 101c - 101c.

[32] Steyvers, M., and Griffiths, T., Probabilistic Topic Models, Lawrence Erlbaum, 2007.

[33] http://www.hsrd.research.va.gov/for_researchers/vinci, accessed Sept. 14, 2012.

[34] Yang, L., Mei, Q., Zheng, K., and Hanauer, D.A., "Query Log Analysis of an Electronic Health Record Search Engine", AMIA Annu Symp Proc, 2011(2011, pp. 915-924.

[35] Zeng, Q.T., Crowell, J., Plovnick, R.M., Kim, E., Ngo, L., and Dibble, E., "Assisting Consumer Health Information Retrieval with Query Recommendations", J Am Med Inform Assoc, 13(1), 2006, pp. 80-90.

[36] Zeng, Q.T., Redd, D., Rindflesch, T., and Nebeker, J., "Synonym, Topic Model and Predicate-Based Query Expansion for Retrieving Clinical Documents", AMIA 2012 Annual Symposium, (In Press)

CHAPTER 3

MAXIMIZING CLINICAL COHORT SIZE USING FREE TEXT QUERIES

Reprinted from Computers in Biology and Medicine, Vol. 60, Gundlapalli AV, Redd D, Gibson BS, Carter M, Korhonen C, Nebeker J, Samore MH, Zeng-Treitler Q, Maximizing clinical cohort size using free text queries, Pages 1-7, Copyright 2015, with permission from Elsevier.

Computers in Biology and Medicine 60 (2015) 1–7

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Maximizing clinical cohort size using free text queries

Adi V. Gundlapalli [a,b,1], Doug Redd [a,c,1], Bryan Smith Gibson [a,b], Marjorie Carter [a,b], Chris Korhonen [a], Jonathan Nebeker [a,b], Matthew H. Samore [a,b,c], Qing Zeng-Treitler [a,c,*]

[a] IDEAS Center, VA Salt Lake City Health Care System, Salt Lake City, UT, USA
[b] Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA
[c] Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, USA

ARTICLE INFO

Article history:
Received 4 September 2014
Accepted 9 January 2015

Keywords:
Gingko
Warfarin
Overweight
Diabetes
Text query
Structured data
Cohort identification
Unstructured data
Clinical notes

ABSTRACT

Background: Cohort identification is important in both population health management and research. In this project we sought to assess the use of text queries for cohort identification. Specifically we sought to determine the incremental value of unstructured data queries when added to structured queries for the purpose of patient cohort identification.
Methods: Three cohort identification tasks were evaluated: identification of individuals taking gingko biloba and warfarin simultaneously (Gingko/Warfarin), individuals who were overweight, and individuals with uncontrolled diabetes (UCD). We assessed the increase in cohort size when unstructured data queries were added to structured data queries. The positive predictive value of unstructured data queries was assessed by manual chart review of a random sample of 500 patients.
Results: For Gingko/Warfarin, text query increased the cohort size from 9 to 28,924 over the cohort identified by query of pharmacy data only. For the weight-related tasks, text search increased the cohort by 5–29% compared to the cohort identified by query of the vitals table. For the UCD task, text query increased the cohort size by 2–43% compared to the cohort identified by query of laboratory results or ICD codes. The positive predictive values for text searches were 52% for Gingko/Warfarin, 19–94% for the weight cohort and 44% for UCD.
Discussion: This project demonstrates the value and limitation of free text queries in patient cohort identification from large data sets. The clinical domain and prevalence of the inclusion and exclusion criteria in the patient population influence the utility and yield of this approach.
Published by Elsevier Ltd.

1. Introduction

Recently healthcare has come to embrace the potential of 'analytics' on large datasets for operations, quality improvement and clinical and biomedical research [1]. With the increase in the availability of large healthcare datasets, there are significant challenges in accessing and making use of these data. There is a need for tools and protocols to harness and 'tame' these big data so that they can be converted to information and ultimately into actionable knowledge. Otherwise, there is a real danger of amassing large datasets that never fulfill the potential of improving patient care and increasing healthcare efficiency.

One critical component of big data analysis is cohort identification, i.e. the reliable identification of a group of patients of interest. Cohort identification is typically an iterative process in which the goal is to separate definite 'cases' of interest from possible cases and then find appropriate controls [2]. Determining whether a given patient meets the definition of a case is often challenging and requires accurate and reliable data. Current database search capabilities support cohort identification within electronic medical record (EMR) systems, however, these are limited to cohort identification based on structured data [3–8]. Recently, there is increasing interest in the use of clinical notes for cohort identification [9–14].

In this study, we sought to determine the value of adding free text search to structured data queries as compared to structured data query alone for cohort identification from an extremely large dataset (> 17 million unique individuals). The study employed Voogo, a user-friendly clinical data search engine that executes key word searches of clinical free text. We compared the cohorts identified using free text search to structured-data-only queries, and did so in three cases of increasingly complex inclusion criteria. A random sample of electronic medical notes from 500 patients was then evaluated by human review to estimate the positive predictive value of correctly identifying cases using free text queries. Our overall goal was to determine the value of free text

* Correspondence to: Mail Code 182, VA Salt Lake City Health Care System, 500 Foothill Drive, Salt Lake City, UT 84148, USA. Tel.: +1 801 213 3357; fax: +1 801 581 4297.
E-mail address: q.t.zeng@utah.edu (Q. Zeng-Treitler).
[1] Equal contribution.

http://dx.doi.org/10.1016/j.compbiomed.2015.01.008
0010-4825/Published by Elsevier Ltd.

queries in cohort identification by answering the question: how many additional patients and observations can be identified through free-text queries as compared to using structured data queries only?

1.1. Clinical setting

The Veterans Health Administration (VHA) operates one of the largest integrated healthcare systems in the United States. The VHA has seen a steady increase in enrollment in recent years with 8.6 million total living enrollees in fiscal year 2012. With regard to volume of patient care, in 2012 there were 79.8 million outpatient visits across VHA's hospital-based clinics and 827 community based outpatient clinics, and VHA's 151 hospitals handled more than 692000 inpatient admissions [16].

1.2. Data corpus

The VA was at the forefront of the development of Electronic Health Records with its nationwide implementation of VistA in 1996. Therefore, VA now has extensive longitudinal records on its millions of enrolled veterans.

Recognizing the opportunities for research using this aggregated data, the VA Health Services Research and Development (HSR&D) office funded the Veterans Informatics and Computing Infrastructure (VINCI), which began operations in June 2008 [15]. VINCI is a service-level collaboration between the Office of Information and Technology (OI&T) and the Office of Research and Development (OR&D), designed to serve the data and IT needs of the VA research community. VINCI provides centralized access to VA data resources in a high-performance computing environment with secure access to comprehensive VA healthcare data. VINCI's mission is to provide researchers with an environment for efficient, secure analysis of patient level data, and to provide tools and coordination for research in basic and applied medical informatics.

The VINCI database is hosted on a Microsoft SQL Server DBMS installation. It is extremely large, at the time of this study providing access to structured and unstructured electronic medical data on 17,543,172 unique patients, living or deceased, since Oct. 1999. VINCI data is updated from raw VistA electronic health records on a nightly basis, and also provides snapshots at the end of the fiscal year [16]. The document corpus consists of 2,096,957,070 clinical documents from providers. The dataset also includes 1,611,284,360 diagnostic codes (ICD9), data on 1,654,598,048 pharmacy prescriptions, and 5,856,426,293 lab tests (both orders and results). Many other types of administrative and clinical data are also available for exploration and discovery. In order to access data, researchers request a VINCI workspace and submit a data access request form after acquiring approval from the institutional review board. The data request indicates the criteria and domains being requested. If approved, data managers provide access to the approved data set.

2. Methods

2.1. Selection of cohort identification tasks

The use cases for this study were derived from real world inquiries we received from clinicians and researchers across the VA system during the first half of 2013. They are representative of the progressively complex cohorts requested by researchers in a US large healthcare system.

The first cohort identification task involves identifying patients who are concurrently taking the herbal remedy *Ginkgo biloba* and the anticoagulant warfarin. There seems be a high potential for patients to take these two substances simultaneously: a recent article reports that 49% of Americans use dietary supplements [17], with *Ginkgo* being a popular herbal supplement. In addition warfarin is a commonly prescribed anticoagulant. The interaction between *Ginkgo* and warfarin is considered to be potentially severe [18]. This cohort identification task therefore represents an attempt to address the important question of the safety of a potentially common drug–supplement interaction.

Though ICD-9-CM codes exist for adverse drug reactions [19], currently, there are no codes for identifying the concurrent usage of *Ginkgo* and warfarin or their interactions. Thus, for this cohort identification task, the only available data source for identifying patients taking *Ginkgo* concurrently with warfarin is through text data as recorded by the medical provider. *Ginkgo* was noted to be dispensed by certain pharmacies in the VA system, thus it was possible to identify prescriptions through the pharmacy databases. Table 1 shows the availability of data elements in the VA electronic medical record to identify this cohort of patients. We used all clinical notes, which include over 2000 note types. The most frequent included "Nursing Note", "Primary Care Note", "Telephone Encounter Note", "Nursing Inpatient Note", and "Primary Care Outpatient Note" [20].

The second cohort identification task involved classifying individuals as either overweight or obese based on their recorded measurements such as weight, height and abdominal girth (or waist circumference). While the majority of evidence has pointed to a clear linear relationship between weight and health, a recent report found a 6% reduction in morbidity among overweight individuals as compared to normal weight individuals (the "obesity paradox") [21]. The seemingly complex association between weight and health is complicated by the many potential effect modifiers: including physical fitness [22], physical activity [22,23], visceral vs. subcutaneous fat depots [24], and genetics [23]. Any study that sets out to examine these complex relationships will need to first classify individuals by their weight status and then account for the time varying nature of this measure (e.g. describe the weight trajectories of individuals in the cohort). This cohort identification task therefore represents an attempt to both capture additional individuals by including free text data but also to increase the number of measures within individuals with which to improve the classification of their weight trajectory.

Table 1
Availability of data elements in the VA electronic medical record to identify specific cohorts of patients.

	Structured data					Unstructured data
	ICD-9-CM codes	Problem lists	Vital statistics table	Laboratory databases	Pharmacy databases	Clinical notes
Concurrent use of *Ginkgo* and warfarin	No	No	No	No	Yes	Yes
Patient weight, height and abdominal girth	No	No	Yes	No	No	Yes
Uncontrolled diabetes	Yes	Yes	No	Yes	No	Yes

In principle, height, weight, and abdominal girth should always be captured as structured data elements and should be available in the vital statistics table. However several reports using VA structured data have found a significant percentage of individuals are missing these data. Littman et al. reported that among the records of 173,127 veterans, 32.8% had missing data for weight or height [25]. Similarly, Das et al. reported that among 1.8 million veterans who received outpatient care at VA facilities in 2000, 50.4% had no recorded height or weight as structured data [26]. This situation is likely to improve as the VA seeks meaningful use certification [27]. It appears that some of the missing data is the result of clinicians entering these measurements as text in their notes rather than structured data in the EHR (personal communication Ken Jones, VA National Program Director for Weight Management, June 30, 2013). This data occurs in narrative discussions as well as in semi-structured sections of notes, such as those labeled Physical Exam (PE), Review of Systems (ROS), etc., however these sections are not consistently named and do not have a consistent structure. Therefore to define this cohort we sought to supplement the structured data in the vitals table with extraction of free text mentions of these measures in the medical note.

The third cohort identification task involved identifying patients whose diabetes was uncontrolled. Diabetes is a common condition among veterans served by VHA [28], and it is of the utmost urgency that the factors associated with lack of control are identified to inform strategies and policies to improve control. While some of these factors are known in the general population, it is important that we study these in specific veteran populations that are particularly at high risk of the complications of uncontrolled diabetes.

Several structured data sources are readily available for assembling this cohort: there are ICD-9-CM codes for uncontrolled diabetes, and a query of the laboratory database would identify patients whose hemoglobin A1C (HbA1C) is $>7.0\%$ indicating poorly controlled diabetes. We hypothesized that the free text in the clinical note would provide additional evidence for uncontrolled diabetes and help identify veterans who did not have relevant ICD codes and lab values in the VA system. This is possible because VA patients may receive care from outside VA either on a routine or emergent basis [29].

These three uses are a convenience sample representing three different contexts of cohort identification: (1) Gingko/Warfarin – we have prior knowledge of the low prevalence of Gingko in coded data; (2) overweight – we have prior knowledge of the limitation of coded data, though coded data do exist; (3) UCD – we have prior knowledge of the limitation of one kind of coded data (ICD) but not that other kind (HbA1C). In all three cases, there is a need to explore the added value of text.

2.2. Voogo: an interactive tool to explore free text and structured data

Voogo is a search engine developed in house by our research group specifically to query VINCI data [30,31]. It supports both free text and structured data searches and provides document, patient, and population-level results (Fig. 1). Searchable fields include document text, document type, age, gender, county, state, Veterans Integrated Service Network (VISN), ICD-9, medication, Current Procedural Terminology (CPT) code, and deceased status. Both Boolean operators and wildcards are supported in query criteria. Query expansion, using synonyms from several UMLS sources (i.e. SNOMED, MeSH, and ICD) and lexical variants from the SPECIALIST lexicon, is also implemented [30]. Results can be saved as lists of patient and document identifiers, or complete result summaries with sample documents and include the geographical distribution and detailed patient and document views of results (Fig. 1). Other query summaries include age distribution, living/deceased, gender, and prescribed VA medications. Voogo is currently configured to either directly query the VINCI database tables or query through the Solr/Lucene search engine (Fig. 2) [32].

Prior to the initiation of Voogo development, other tools were explored that were available or could be made available in the VINCI environment. None were found that met our criteria of a graphical user interface for search construction on both structured and unstructured data; integration with the existing VA corporate data warehouse; allowance for integration of new, flexible indexing; result visualization; flexible result sampling; and a query expansion function. Voogo takes advantage of full-text search features of Microsoft SQL Server as well as the Solr/Lucene search platform. No suitable search tools were found that supported negation assertions. Voogo is not an in depth NLP tool like V3NLP, HITEx, Sophia, cTAKES, etc. which are part of eMERGE, CHIR, and SHARP [33–36]. Rather it is intended for text and coded data exploration and cohort identification. Negation assertion is planned for the near future. Voogo is open source software in the process of being released through the VA/OSEHRA mechanism.

2.3. Free text and structured data queries to identify cohorts of patients

The research team developed structured and free-text queries for each of the cohort identification tasks. The queries were iteratively revised after reviewing interim results until the researchers were satisfied that an appropriate population was identified. For example, based on initial results the queries for *Ginkgo biloba* use were revised to include two spelling variants: "*Gingko*" and "*Ginko*". Table 2 shows the final queries used in the study. Ultimately, researchers who request cohorts determine the termination point in query formation. During the iterative process, the investigators are discovering new query terms and assess the usefulness of terms for the tasks, in an informal fashion. While standard vocabularies exist (and are taken advantage of by our query recommendation service), the vocabularies are far from perfect and clinical judgments are required.

2.4. Manual review to determine positive predictive values of retrieved results

For each of the three cohort identification tasks, a random sample of all clinical notes for 500 veterans (100 each for *Gingko*, height, weight, girth, and uncontrolled diabetes) was drawn from the full set of documents identified by the free text queries. A total of 6425 documents were extracted for human review (*Ginkgo*=433, weight=3043, height=2086, girth=356, UCD=507). These records were reviewed to determine whether free text queries identified (1) true positives (TP) or false positives (FPs); (2) the same patients as structured data queries. The records were raw, not de-identified; the local institutional review board (IRB) approved all data access and use in this study, and no data was accessed or shared without prior IRB approval.

The positive predictive values (PPVs) of using text for case identification were calculated based on chart reviews. Experienced reviewers developed guidelines to establish true positive cases for each of the three use cases. For *Gingko* and warfarin usage, we relied on chart review alone to determine if *Gingko* was actively being used (or had been discontinued). In the weight-related use case, manual review determined the presence of a specific value for weight, height, and abdominal girth. For the review of charts related to diabetes, we deemed the record a positive if there was at least one instance of mention of uncontrolled diabetes based on the terms used for the free text query. Two experienced researchers reviewed a random sample of query search results. The inter-reviewer agreement was calculated for

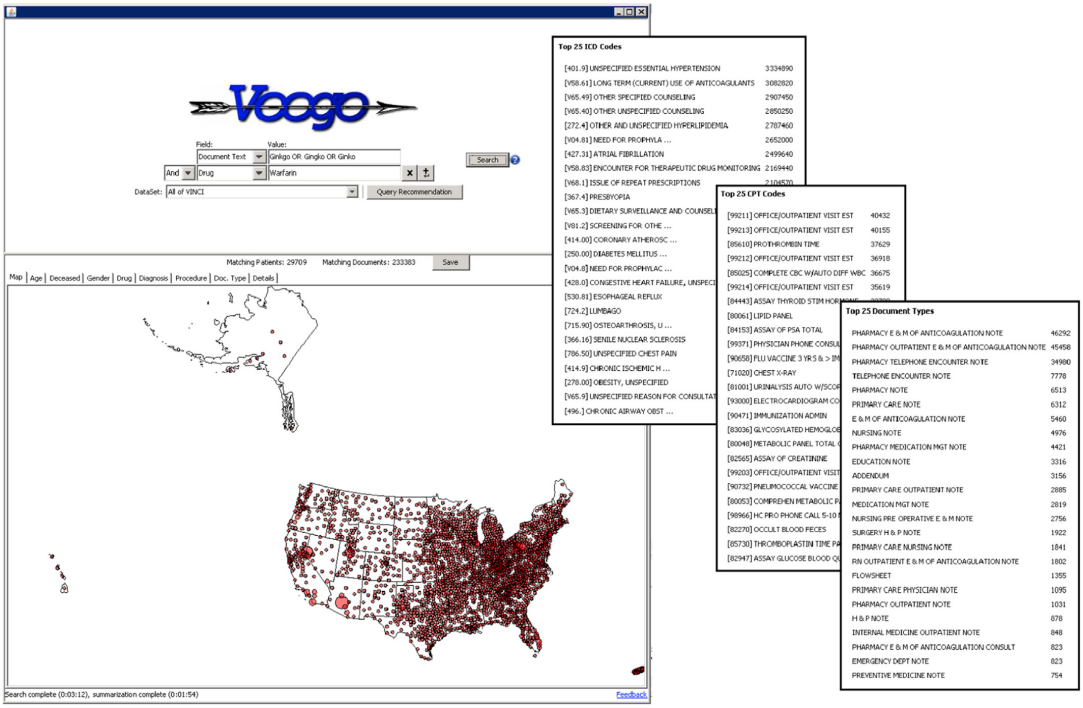

Fig. 1. Screen shot of query results in Voogo, diagnosis, procedure, and document type views.

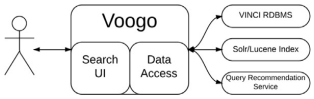

Fig. 2. Overview of Voogo architecture. The VINCI RDBMS is the primary database data source for EHR data, with the Solr/Lucene Index providing enhanced text search features. The Query Recommendation Service assists querying by suggesting additional related terms to include in the query.

each chart review; discordances were adjudicated by discussion among the reviewers until consensus was reached.

The overall workflow for the cases is (1) Use Voogo to iteratively preview data and refine query; (2) perform query and assess potential value in addition to structured data; (3) if free text adds value, obtain cohort and assess specificity; (4) if specificity is not satisfactory, train NLP to refine cohort; (5) analyze data, using NLP results in the analysis if necessary (e.g. the weight value). The general use of text to maximize cohort size is applicable to other EHR systems, although the specific Voogo tool is designed for the VINCI system.

2.5. NLP of ginkgo and weight notes

In order to compare query results with what can be accomplished with additional NLP, we performed NLP analysis of *Gingko* and weight notes. Using the 433 cases that had been manually reviewed for *Gingko*, we first manfully crafted a set of processing rules to highly prevalent templates. We then trained a support vector machine (SVM) model using the notes not covered by the template rules. The SVM developed was conducted using WEKA SMO algorithm along the default parameters. The final NLP module first applies the template rules and then applies the SVM model. We tested the NLP module on 200 randomly selected notes retrieved by the Gingko query and calculated accuracy measures. Because the evaluation results are good, the NLP module was then applied to all retrieved notes to filter out FP cases. We also developed an NLP module for extraction of weight values from free text notes retrieved by the weight query. We used a novel Regular Expression Discovery (RED) algorithm to automatically discover regular expressions to extract weight values. We trained the RED Extraction model using 571 manually reviewed primary care outpatient notes, and applied the model to 5716 notes of from 1000 random patients.

3. Results

3.1. Identification of cohorts

Large numbers of patients were identified for all three cohorts (Table 3). In most cases, text searches identified more patients than structured data queries alone. Using structured-data alone, only nine patients were found to be taking both *Gingko* and warfarin. Free text queries returned over 28,000 patients. Free text queries returned roughly the same number of patients for weight and height as structured data queries. More patients with girth information or diabetes complications were identified through text queries.

3.2. Precision and recall for the three use cases

There were a total of 6425 documents reviewed for the 500 patients randomly selected for the three cohort identification tasks (100 each for *Gingko*, height, weight, girth, and uncontrolled diabetes).

Table 2
Description of manually curated queries using structured data and free text notes to identify specific cohorts of patients from the VA electronic medical record.

Concurrent use of *Ginkgo* & Warfarin	
***Ginkgo* (structured data)**	Queried the filled prescriptions table for *Ginkgo* and variants "gingko" and "ginko". There were no occurrences of the variants, so future queries on this table only used *Ginkgo*.
***Ginkgo* (free text notes)**	Queried the clinical documents table for documents containing *Ginkgo* and variants "gingko" and "ginko". Of the matching documents approximately 50% used *Ginkgo*, 25% "gingko", and 25% "ginko".
Warfarin	Queried the filled prescriptions table for warfarin and its alternate brand names (Jantoven, Coumadin, Marevan, Lawarin, Waran, and Warfant). There were no occurrences of the alternate brand names, so future queries only used warfarin.
Patient weight, height and abdominal girth	
Structured data	Queried the vital statistics table for measurements of weight, height, and circumference/girth or abdominal girth.
Free text notes	Queried the clinical documents table for documents containing "weight" OR "wt", "height" OR "ht" AND "girth" or "waist circumference".
Uncontrolled diabetes	
Structured data	Queried patient diagnosis table for ICD 9 codes for uncontrolled diabetes, including 250.02, 250.03, 250.12, 250.13, 250.22, 250.23, 250.32, 250.33, 250.42, 250.43, 250.52, 250.53, 250.62, 20.63, 250.72, 250.73, 250.82, 250.83, 250.92, and 250.93.
Free text notes	Queried the clinical documents table for documents containing text indicating uncontrolled diabetes, including "diabetes" in combination with "uncontrolled", "out of control", "lack of control", "poor control", "not well controlled", "poor compliance", "non compliant", "high sugars", "high glucose levels", "high A1C", "high HbA1C", "hyperglycemia", or "end stage diabetes".

Table 3
Comparison of numbers of patients identifiable from structured data and free text notes, and number of patients identifiable from both sources.

	# Patients (structured data)	# Patients (free text notes)	# Overlapping patients	% Overlapping patients
***Ginkgo*+Warfarin**	9	28,924	8	0.03
Weight, height and abdominal girth				
Weight	9,275,267	9,686,634	8,811,207	86.80
Height	9,064,078	8,985,076	8,274,961	84.66
Girth	510,934	931,312	196,293	15.75
Uncontrolled diabetes				
ICD-9-CM	758,069	2,136,615	614,463	26.95
HbA1C 7–9	1,514,906		1,062,717	41.05
HbA1C > 9	745,398		615,741	27.17
HbA1C < 7	1,613,194		1,116,634	42.41

Inter-rater agreement for the reviews and positive predictive values for each of the queries are presented in Table 4. The highest PPVs were found for height and weight, while the lowest PPV was associated with girth. Structured data were poor in identifying true positives for *Ginkgo*+warfarin and abdominal girth. The increase in cohort size was dramatic for *Gingko* and warfarin (167,116 fold increase), and of practical significance for abdominal girth (29%). It was interesting to note that even for situations where structured data exist (ICD codes and laboratory results for uncontrolled diabetes), there was a significant increase in cohort size. This was likely due to inconsistent capture of structured data. Table 5 provides examples of true positives (TPs) and false positives (FPs) found in these cohort identification tasks.

In the use case involving patient weight and height, free text queries essentially doubled the number of observations (Table 6). However, most of the cases from free text queries were also identified by structured data.

3.3. NLP of ginkgo and weight notes

On the 200 randomly selected *Gingko* notes, the NLP model reached a PPV of 90%, sensitivity of 97%, specificity of 78%, and *F* measure of 93%. 10-fold cross-validation of the Weight model gave a PPV of 98.8%, sensitivity of 98.3%, specificity of 98.1%, and *F* measure of 98.5%. Comparison to weights from structured vital signs data showed that 7.7% of the weight measurements from text were not available in the structured data.

4. Discussion

In this study we examined the incremental value of free text queries for identification of cohorts of patients from a very large clinical dataset. Our results suggest that free text queries on big clinical data can add value to the task when compared with searches using structured data alone, especially in cases where structured data do not exist. In all three-cohort identification tasks, text queries increased the size of patient cohorts. In some cases, the percentage of increase was relatively small (8% of the true positive patients identified by height text query had no height recorded in vital signs). While in other cases the increase was truly dramatic: virtually all of the true positive concurrent *Gingko*-warfarin users were identified by text queries. Finally, tracking patient weight using both structured data and free text notes not only identified more patients but also increased the number of observations per patient. This suggests that certain clinical variables are more consistently recorded in both text and structured data (e.g. height and weight), while text might be the only source for other data such as *Ginkgo biloba* use. When high quality structured data are available as in the case of HbA1C, the added value of text query is clearly more limited. However, we observe that not all structured data are of equal quality. Using ICD codes to identify uncontrolled diabetes, for instance, failed to identify many patients.

Simply stated, we found that free text queries can be fruitfully combined with structured data search to yield a more complete patient cohort from the electronic medical record. In looking at the different use cases, the utility of free text queries varies by the clinical domain and the prevalence of the inclusion and exclusion criteria in the patient population. Furthermore, the key to reliable and complete patient identification is the availability of individual data elements in the EMR and the ability to access them.

We acknowledge several limitations. In this study, the precision and recall of text queries varied by the query. High precisions were observed for height and weight. Lower rates were found for *Gingko*, uncontrolled diabetes, and abdominal girth. Through our experience in using this tool we have developed several strategies to filter out false positives (FPs). In certain queries, the FPs or TPs are concentrated in specific document types or notes from certain facilities.

Table 4
Results of manual review of patient medical notes: positive predictive value, inter-rater agreement, and estimated increase of cohort size from free text queries.

	Positive predictive value (precision)[a]	Sensitivity (recall)[b]	Cohen's Kappa for manual review (inter-rater agreement)	% True positives not identified in structured data (%)	Estimated increase of cohort size using free text queries *N* (%)
***Ginkgo*+Warfarin**	0.52	1.00	0.82	100[c]	15,040 (167,116%)
Data related to obesity					
Weight	0.94	0.95	0.90	8	728,435 (8%)
Height	0.96	0.92	1.00	5	431,284 (5%)
Girth	0.19	0.36	0.73	84	148,637 (29%)
Data related to uncontrolled diabetes					
ICD	0.44	0.87	0.87	35	329,039 (43%)
HbA1C 7–9		0.68		6	56,407 (4%)
HbA1C > 9		0.88		50	470,055 (63%)
HbA1C < 7		0.6		3	28,203 (2%)

[a] True positives determined from manual review.
[b] False negatives determined from patients identified using structured data but not identified using free-text notes.
[c] 99.99…%, rounded to 100%.

Table 5
Random examples of true positives and false positives as determined by manual review of free text notes returned from free text queries.

	True positive	False positive
***Ginkgo*+Warfarin**	MEDS: *GINKGO BILOBA*, ECHINACEA, FISH OIL	Patient decided not to take it (*Gingko biloba*)
Patient measurements		
Weight	Weight: 178 lb. [80.7 kg]	No unwanted weight loss
Height	Height: 68.5 in	Adjusted the height of the rollator
Abdominal girth	Waist circumference: 47 in.	No data available for abdominal girth
Uncontrolled diabetes	Diabetes is not well controlled	No diabetes, poor control of hypertension

Table 6
Comparison of number of observations identifiable from structured data and free text notes for a selection of patients with both types of data.

	True positive overlapping patients from manual review sets		
	# Patients	# Observations structured data	# Observations free text notes
Height	92	727	1775
Weight	89	1511	2988

Determining these sources of error and applying appropriate filters has the potential to reduce false positives. For instance, many of the *Gingko* FPs were noted to be in pre- and post-operative instructions to patients. Excluding those types of documents could improve the PPV. Complex natural language processing (NLP) may be required to filter out other source of FPs such as negation, templating within the notes, hypotheticals (e.g. if you have high blood sugars, then increase your insulin), clinical plans, instructions (e.g. please stop your warfarin at least 7 days prior to your surgery), family history, and past history. In evaluating additional NLP on both *Gingko* and weight we showed that NLP is able to increase accuracy measures, more so with *Gingko* than weight, demonstrating that NLP is necessary where higher accuracy measures are required, although at much higher cost.

Our chart review was limited to only 100 randomly selected patients for each text query (500 patients total); even though this included 6425 documents, the results of our manual review may not be representative of the full document corpus. With several million patients in each cohort and several hundred million documents, a human review of even a meaningful sample is a daunting task. The human review effort utilized approximately 150 person hours for the 6425 documents from 500 patients. By simple extrapolation, manual annotation of these features for the over 17 million patients in the entire data corpus would equate to roughly 5,000,000 person hours, or over 500 person years. The chart review was conducted on positively identified cases, since the number of negatives cases is very large and the prevalence of any given condition is low. This makes estimating the recall or sensitivity particularly challenging when working with such a large dataset, and points to the pressing need to develop better methods of evaluating results of analyses of big data. Since we cannot fully assess false negatives (FNs), text search generally casts a wide net. In our use cases the queries are the product of iterative search and revision to minimize FNs at the cost of higher FPs. Although more FPs increases the NLP burden, NLP can correct for this. Loss of TPs at the stage of cohort identification cannot be reversed by NLP. At each step of the process, quality evaluation is important as errors can be accumulated and passed on to subsequent steps in the process.

In conclusion, this study demonstrates the power of adding free text queries to the task of cohort identification using a large clinical dataset. The result is the 'taming' of these big data to a manageable size. Using this efficient free text query approach requires minimal human and computing resources and may be the endpoint for some cohort identification tasks, while for more detailed and sophisticated projects it may serve as a starting point.

Conflict of interest statement

None declared.

Acknowledgments

Funding for this project was provided by U.S. Department of Veterans Affairs, Veterans Health Administration HSR and D, Office of Research and Development, Health Services Research and Development Projects CHIR HIR 08-374, VINCI HIR-08-204 and HIR 10-002. Further support was provided by NIH grants 1R01LM011334

and 1R01AT006548-01A1. Resources and administrative support were provided by the VA Salt Lake City Health Care System (IDEAS Center). We thank our various team members for their assistance with this project. We appreciate and acknowledge our colleagues at VA Informatics and Computing Infrastructure (VINCI) for their assistance with accessing VA 'big data'. The views expressed in this paper are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs or the United States Government.

References

[1] T.B. Murdoch, A.S. Detsky, The inevitable application of big data to health care, J. Am. Med. Assoc. 309 (13) (2013) 1351–1352.
[2] J.C. Denny, Chapter 13: Mining electronic health records in the genomics era, PLoS Comput. Biol. 8 (12) (2012) e1002823.
[3] M. Cuggia, et al., Roogle: an information retrieval engine for clinical data warehouse, Stud. Health Technol. Inf. 169 (2011) 584–588.
[4] S.B. Martins, et al., Evaluation of KNAVE-II: a tool for intelligent query and exploration of patient data, Stud. Health Technol. Inf. 107 (Pt 1) (2004) 648–652.
[5] M.M. Horvath, et al., The DEDUCE Guided Query tool: providing simplified access to clinical data for research and quality improvement, J. Biomed. Inf. 44 (2) (2011) 266–276.
[6] D.A. Hanauer, EMERSE: the electronic medical record search engine, AMIA Annu. Symp. Proc. (2006) 941.
[7] L. Seyfried, et al., Enhanced identification of eligibility for depression research using an electronic medical record search engine, Int. J. Med. Inf. 78 (12) (2009) e13–e18.
[8] L. Yang, et al., Query log analysis of an electronic health record search engine, AMIA Annu. Symp. Proc. 2011 (2011) 915–924.
[9] A.A. Okan Ozturkmenoglu, DEMIR at TREC-Medical 2011: Power of Term Phrases in Medical Text Retrieval, in: Proceedings of 20th Anniversary of Text Retrieval Conference, 2011.
[10] J. Chung, S. Murphy, Concept–value pair extraction from semi-structured clinical narrative: a case study using echocardiogram reports, AMIA Annu. Symp. Proc. (2005) 131–135.
[11] D.K. Mariam Daoud, Jun Miao, Jimmy Huang, York University at TREC 2011: Medical Records Track, in: TREC 2011. Gaithersburg, Maryland, USA, 2011.
[12] D.T. Martijn Schuemie, Edgar Meij, DutchHatTrick: Semantic query modeling, ConText, section detection, and match score maximization, in: TREC 2011, 2011.
[13] S.K.D.M.S. Ghodke, , L.Z.H. Suominen, L. Cavedon, Search for Medical Records: NICTA at TREC 2011 Medical Track, in: TREC 2011, 2011.
[14] L.T. Duy Dinh, IRIT at TREC 2011: evaluation of query expansion techniques for medical record retrieval, in: Text Retrieval Conference, TREC 2011. Gaithersburg, Maryland, USA.
[15] US Department of Veterans Affairs. VA Informatics and Computing Infrastructure (VINCI), cited 2013. Available from: ⟨http://www.hsrd.research.va.gov/for_researchers/vinci/⟩, 2013.
[16] VA Information Resource Center (VIReC). VHA Corporate Data Warehouse (CDW), cited. Available from: ⟨http://vaww.virecresearch.va.gov/CDW/Overview.htm⟩, 2014.
[17] R.L. Bailey, et al., Dietary supplement use in the United States, 2003–2006, J. Nutr. 141 (2) (2011) 261–266.
[18] K.M. Bone, Potential interaction of Ginkgo biloba leaf with antiplatelet or anticoagulant drugs: what is the evidence? Mol. Nutr. Food Res. 52 (7) (2008) 764–771.
[19] P. Hougland, et al., Performance of International Classification Of Diseases, 9th Revision, Clinical Modification codes as an adverse drug event surveillance system, Med. Care 44 (7) (2006) 629–636.
[20] Q.T. Zeng, et al., Characterizing clinical text and sublanguage: a case study of the VA clinical notes, J. Health Med. Inf. 4 (2011) 2.
[21] K.M. Flegal, et al., Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis, J. Am. Med. Assoc. 309 (1) (2013) 71–82.
[22] G.E. Duncan, The fit but fat concept revisited: population-based estimates using NHANES, Int. J. Behav. Nutr. Phys. Act. (7) (2010) 47.
[23] C.S. D'Angelo, et al., Obesity with associated developmental delay and/or learning disability in patients exhibiting additional features: report of novel pathogenic copy number variants, Am. J. Med. Genet. A 161 (3) (2013) 479–486.
[24] H.E. Bays, et al., Pathogenic potential of adipose tissue and metabolic consequences of adipocyte hypertrophy and increased visceral adiposity, Expert Rev. Cardiovasc. Ther. 6 (3) (2008) 343–368.
[25] A.J. Littman, et al., Evaluation of a weight management program for veterans, Prev. Chronic Dis. 9 (2012) E99.
[26] S.R. Das, et al., Obesity prevalence among veterans at Veterans Affairs medical facilities, Am. J. Prev. Med. 28 (3) (2005) 291–294.
[27] M. Mosquera, VA Unveils Plans to Certify VistA for Meaningful Use. Available from: ⟨http://www.healthcareitnews.com/news/va-unveils-plans-certify-vista-meaningful-use⟩, 2012 (01.12.14).
[28] D.R. Miller, M.M. Safford, L.M. Pogach, Who has diabetes? Best estimates of diabetes prevalence in the Department of Veterans Affairs based on computerized patient data, Diabetes Care 27 (Suppl. 2) (2004) B10–B21.
[29] M. Ajmera, T.L. Wilkins, U. Sambamoorthi, Dual Medicare and Veteran Health Administration use and ambulatory care sensitive hospitalizations, J. Gen. Intern. Med. 26 (Suppl. 2) (2011) S669–S675.
[30] Q.T. Zeng, et al., Synonym, topic model and predicate-based query expansion for retrieving clinical documents, AMIA Annu. Symposium Proc. 2012 (2012) 1050–1059.
[31] D. Redd, J. Kuang, Q. Zeng-Treitler, Differences in nationwide cohorts of acupuncture users identified using structured and free text medical records, AMIA Annu. Symp. Proc. (2014) 1002–1009.
[32] Apache Solr. Available from: ⟨http://lucene.apache.org/solr/⟩, November 25, 2014.
[33] Q.T. Zeng, et al., Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system, BMC Med. Inform. Decis. Mak. (2006) 306 (2006) 30.
[34] G. Divita, et al., Sophia: a expedient UMLS concept extraction annotator, AMIA Annu. Symp. Proc. (2014) 467–476.
[35] G.K. Savova, et al., Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (2010) 507–513.
[36] Collaboration Between VINCI and CHIR. Available from: ⟨http://www.hsrd.research.va.gov/for_researchers/vinci/chir.cfm⟩.November 25, 2014.

CHAPTER 4

DIFFERENCES IN NATIONWIDE COHORTS OF ACUPUNCTURE USERS IDENTIFIED USING STRUCTURED AND FREE TEXT MEDICAL RECORDS

Redd D, Kuang J, Zeng-Treitler Q. Differences in nationwide cohorts of acupuncture users identified using structured and free text medical records. AMIA Annual Symposium proceedings / AMIA Symposium. 2014:1002-9. Reprinted with kind permission of The American Medical Informatics Association.

Differences in Nationwide Cohorts of Acupuncture Users Identified Using Structured and Free Text Medical Records

Doug Redd, MS[1,2], Jinqiu Kuang, MS[1], Qing Zeng-Treitler, PhD[1,2]

[1]VA Salt Lake City Health Care System; [2]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah

Abstract

Integrative medicine including complementary and alternative medicine (CAM) has become more available through mainstream health providers. Acupuncture is one of the most widely used CAM therapies, though its efficacy for treating various conditions requires further investigation. To assist with such investigations, we set out to identify acupuncture patient cohorts using a nationwide clinical data repository. Acupuncture patients were identified using both structured data and unstructured free text notes: 44,960 acupuncture patients were identified using structured data consisting of CPT codes;. Using unstructured free text clinical notes, we trained a support vector classifier with 86% accuracy and was able to identify an additional 101,628 acupuncture patients not identified through structured data (a 226% increase). In addition, characteristics of the patients identified through structured and unstructured data were compared, which show differences in geographic locations and medical service usage patterns. Patients identified with structured data displayed a consistently higher use of the Veterans Health Administration (VHA) medical system.

Introduction

Over the past decade, integrative medicine has gained increasing attention from providers and researchers. Compared to traditional healthcare, integrative medicine's emphasis on a partnership between patients and clinicians takes a holistic view of patients' health and well being, and incorporates complementary and alternative medicine (CAM) approaches such as acupuncture and massage into treatment options. Many large hospitals now provide some form of integrative health services to their patients.

At the same time, the safety and effectiveness of many CAM treatments are not sufficiently understood. For instance, acupuncture is widely practiced to relieve pain and treat certain health problems, but debate on its effectiveness continues in the literature. Witt et al evaluated clinical and economical effectiveness of acupuncture on chronic low back pain in a large randomized controlled trial (RCT).(1) They demonstrated that acupuncture in addition to routine care considerably improved clinical outcomes and was relatively cost-effective. A systematic review of RCTs looking at acupuncture for pain was published by Linde et.al. It included thirteen trials (3,025 patients) with a variety of pain conditions and found a small analgesic effect from acupuncture, hardly distinguishable from bias.(2) Another systematic review of 23 RCTs on the effectiveness of acupuncture for nonspecific lower back pain by Yuan et al showed moderate evidence that acupuncture is more effective than no treatment, and strong evidence of no significant difference between acupuncture and sham acupuncture, for short-term pain relief.(3) This review concluded that acupuncture can be a useful supplement to other forms of conventional therapy for nonspecific lower back pain, but the effectiveness of acupuncture compared with conventional therapies requires further investigation. Considering acupuncture is one of the most studied CAM modalities, these uncertain results indicate that more research is needed to ascertain the efficacy of CAM practices.

Secondary analysis of electronic medical records (EMR) is a powerful approach to study treatment safety and effectiveness. At the Veterans Health Administration (VHA), we have begun leveraging its nationwide EMR repository to study the use of acupuncture to manage pain and control other symptoms like nausea. A critical step in EMR secondary analysis is cohort identification.

In this paper, we describe our effort to identify a cohort of patients who had undergone acupuncture treatments while receiving care from the VHA. Both structured data and unstructured data were used. To understand the impact of data source on the resultant cohorts, cohorts identified from the two methods were compared in terms of size of patient characteristics.

Materials and Methods

Data Source

Data for this study was procured through the Veterans Informatics and Computing Infrastructure (VINCI), VHA. The VHA comprises 152 medical facilities in addition to 1,400 clinics that are community-based and tailored to serve individuals on an outpatient basis, Vet Centers, community living centers, and Domiciles. In total, these facilities employ over 53,000 healthcare professionals who provide their services to over 8.3 million veterans on an annual basis. VINCI is a collaboration between the Office of Research and Development and the Office of Information and Technology in the U.S. Department of Veterans Affairs (VA), providing data and infrastructure needs of the VHA research community. VINCI provides access to structured and unstructured health information originating from the VISTA electronic health record system, and includes data for over 17 million patients. We identified patients receiving acupuncture treatments through structured as well as unstructured data using the process outlined in Figure 1.

Cohort Identification Using Structured Data

VHA offers many forms of CAM treatments from acupuncture to sweat lodge. Patients receiving specific treatments within the VHA system can be identified through Current Procedural Terminology (CPT) codes identifying specific patient procedures. Acupuncture treatments are represented by CPT codes 97780, 97781, 97810, 97811, 97813, and 97814.

Many non-standard treatments can be identified through the locations of patient visits. In the VHA, clinic "Stop Codes" are included in the outpatient visit records to indicate the clinic or work group providing specific services. We were, however, only able to identify a single location for acupuncture services using the "Stop Code". Since acupuncture services are widespread in the VHA system, we resorted to CPT codes for their identification.

Cohort Identification Using Free Text Data

Structured data has been shown to be insufficient for cohort identification in many cases (4). Some patients receiving acupuncture will not have corresponding CPT codes assigned for various reasons. For example, many patients obtain treatment from non-VHA providers, particularly when VHA clinics offering a specific therapy are not available in the geographic area of the patient. In some cases, their VHA clinicians do not prescribe or authorize the treatments. Although they may not be recorded by CPT codes, many VHA healthcare providers do ask Veterans about the non-VHA treatments they are receiving and document them in narrative clinical notes. Thus, we searched unstructured, free text clinical notes for mentions of acupuncture.

Free text clinical notes have been shown to be rich in medical information that can be accessed using natural language processing techniques (5-7). Searching of the unstructured notes was accomplished using the Voogo search engine, which was developed specifically for searching structured and unstructured data within VINCI. Using Voogo, patients with clinical documents containing the string "acupuncture" were identified. Snippets of text containing acupuncture, including surrounding context, were extracted and manually annotated to identify if the snippets were positive, negative, or prescribed (if the snippet described a recommendation) for use of acupuncture treatment by the patient. A support vector machine (SVM) was trained for automated acupuncture text classification. Using text classification results, patients were classified as positive for acupuncture treatment use if they had at least one positive snippet; prescribed if they had no positive snippets but at least one prescribed snippet; or negative if they had only negative snippets.
Patients with a positive history of acupuncture use identified through unstructured data is referred to as UD.

Comparing Cohorts from Structured and Free Text Data

We compared the two cohorts (UD and SD) to determine the distribution of patients identifiable only from SD, only from UD, or both. The UD and SD patients were then compared and contrasted for geographic location, gender, age, and most frequent medical procedures, diagnoses, and prescriptions.

A list of the 25 most common procedures was constructed by combining the 21 most frequent Current Procedural Terminology (CPT) codes from UD patients and the 21 most frequent CPT codes from SD patients (the number of codes from UD and SD patients was chosen by trial and error to obtain a combined number of 25). Similarly, a list of the 25 most common diagnoses was constructed by combining the 23 most frequent International Classification of Diseases version 9 (ICD-9) codes from UD and SD patients. And again with prescriptions, the 24 most frequent drug names from UD and SD patients were determined, for a combined set of 25 unique drug names. We then determined the proportions of UD and SD patients receiving these most frequent procedures, diagnoses, and prescriptions..

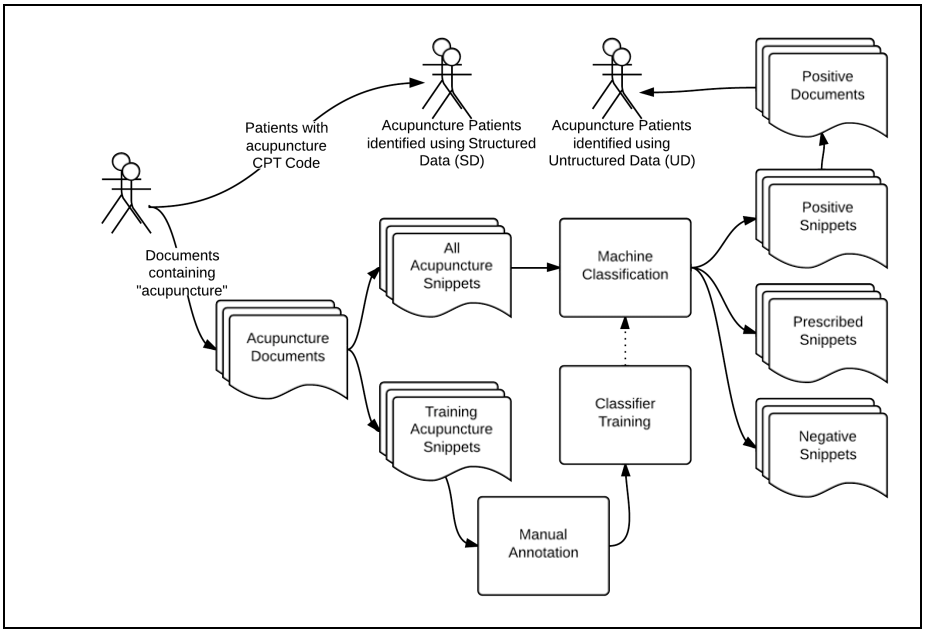

Figure 1. Acupuncture Patient Cohort Identification from Structured Data (SD) and Unstructured Data (UD)

Results

Using CPT codes, 44,960 patients were identified as receiving acupuncture treatment using structured For identification of acupuncture using unstructured data (UD), 1,245,753 documents mentioning identified representing 400,350 patients. 297 snippets were classified as positive, prescribed, or annotators with an inter-rater reliability kappa score of 0.74. Since the kappa is relatively low, reached through discussion to create the reference standard. A support vector machine (SVM) using these snippets and validated with 10-fold cross validation. This resulted in the ability to identify text with an overall accuracy of 0.862 (precision 0.883, recall 0.743, and f_1-measure 0.785) (Table 1). Using the SVM classification model, 140,525 patients were identified as positive for acupuncture use. SD and UD identified patients were compared to determine an intersection of 38,897 patients, so that an additional 101,628 (226%) patients were identified using UD that were not identifiable using SD (Figure 2).

Table 1. Confusion matrix for acupuncture classifier.

		Reference Standard			
		Yes	No		
Acupuncture Classifier	Yes	77	13	Precision	88.3%
	No	24	183	Recall	74.3%
				F1 Measure	78.5%
				Accuracy	86.2%

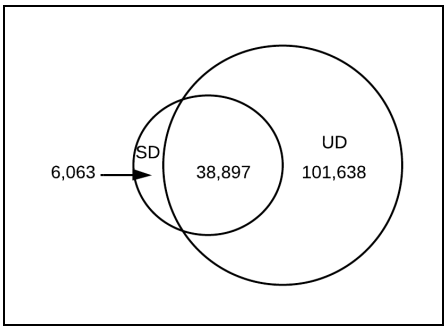

Figure 2. Distribution of patients between groups identifiable by structured data (SD), unstructured data (UD), or both (SD + UD).

We compared the geographic locations of UD and SD patients. Overall, patients congregated around major population centers. There were some differences in the distributions, however. There was a much higher proportion of UD patients in the northwest region of Oregon, and a higher proportion of SD patients in the New York City metropolitan region. (Figure 3).

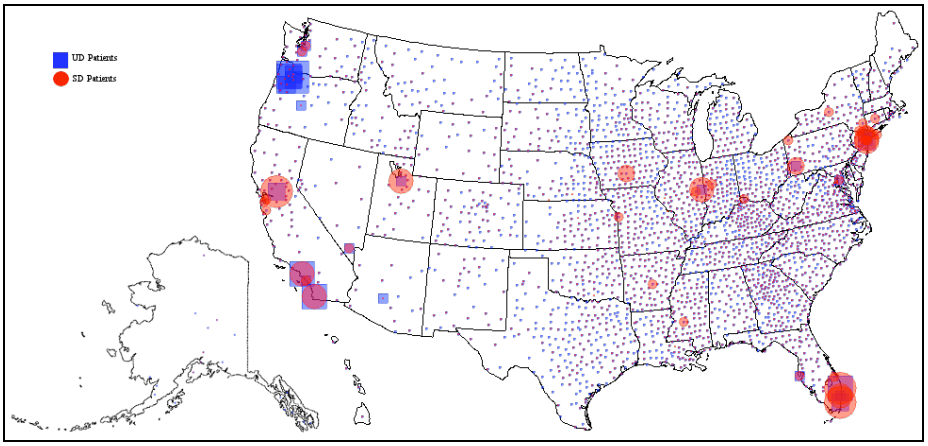

Figure 3. Geographic distribution of acupuncture patients identified from structured data (SD) and unstructured data (UD).

The age distribution of UD and SD patients were essentially similar (Figure 4). The mean age was 56.4 (stdev. 15.2) for UD patients and 56.1 (stdev. 14.6) for SD patients. The difference was statistically significant ($p < 0.00$ by student t-test) due to the large sample size, however this small difference is not clinically meaningful.

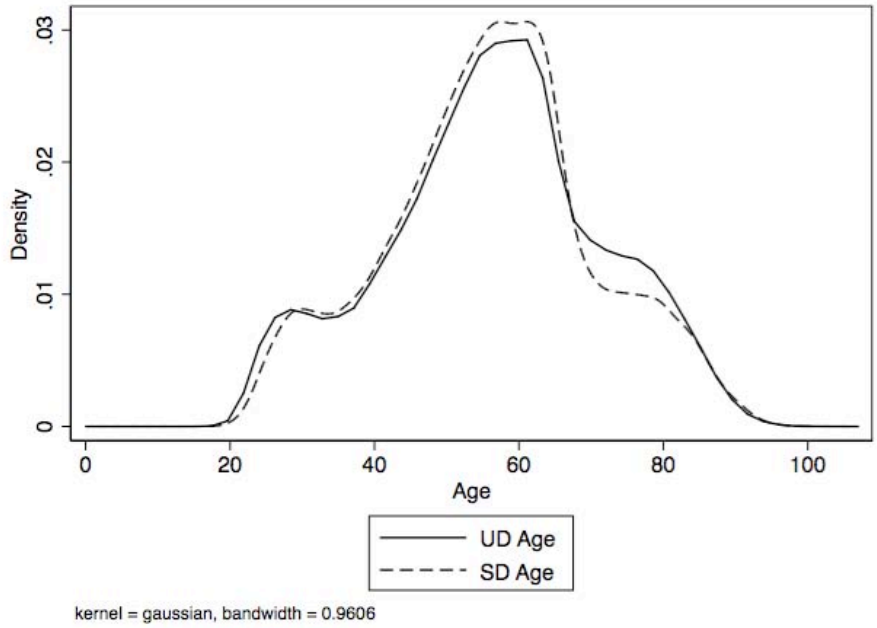

Figure 4. Density of age distribution for acupuncture patients identified from unstructured data (UD) and structured data (SD).

We compared the percent of UD and SD patients receiving the most common procedures, diagnoses, and prescriptions (Figure 5). Although the percentages are somewhat similar, overall a higher percent of SD patients received the measured procedures, diagnoses, and prescriptions. There are some procedures where SD patients show a higher percent that is more pronounced, i.e. metabolic panel total calcium, patient evaluation, therapeutic exercises, and comprehensive metabolic panels. Alternatively, UD patients show a higher percent of assays for quantitative blood glucose, alanine aminotransferase, and assay of urea nitrogen. For diagnoses SD patients also have a higher percentage in most cases, exceptions including unspecified reason for consultation and unspecified tobacco use disorder. Prescriptions continue the trend of higher SD percentages, with SD having higher percentages in all cases. We also examined the per-patient average number of all procedures, diagnoses, prescriptions, and visits between the two groups. This analysis confirmed that SD patients had a higher rate of use in all cases (Table 2).

Overall, both diagnoses and prescriptions indicate the presence of pain and pain management. Diagnoses of lumbago (low back pain), unspecified back pain, and cervicalgia (neck pain) are frequent, as are pain medications such as hydrocodone, gabapentin, cyclobenzaprine, naproxen, tramadol, oxycodone, codeine, etc.

We also compared the gender distribution (Figure 5). There was a higher percent of females in SD patients (14%) as opposed to UD patients (12%), both of which reflect the expected minority of females in the veteran population.

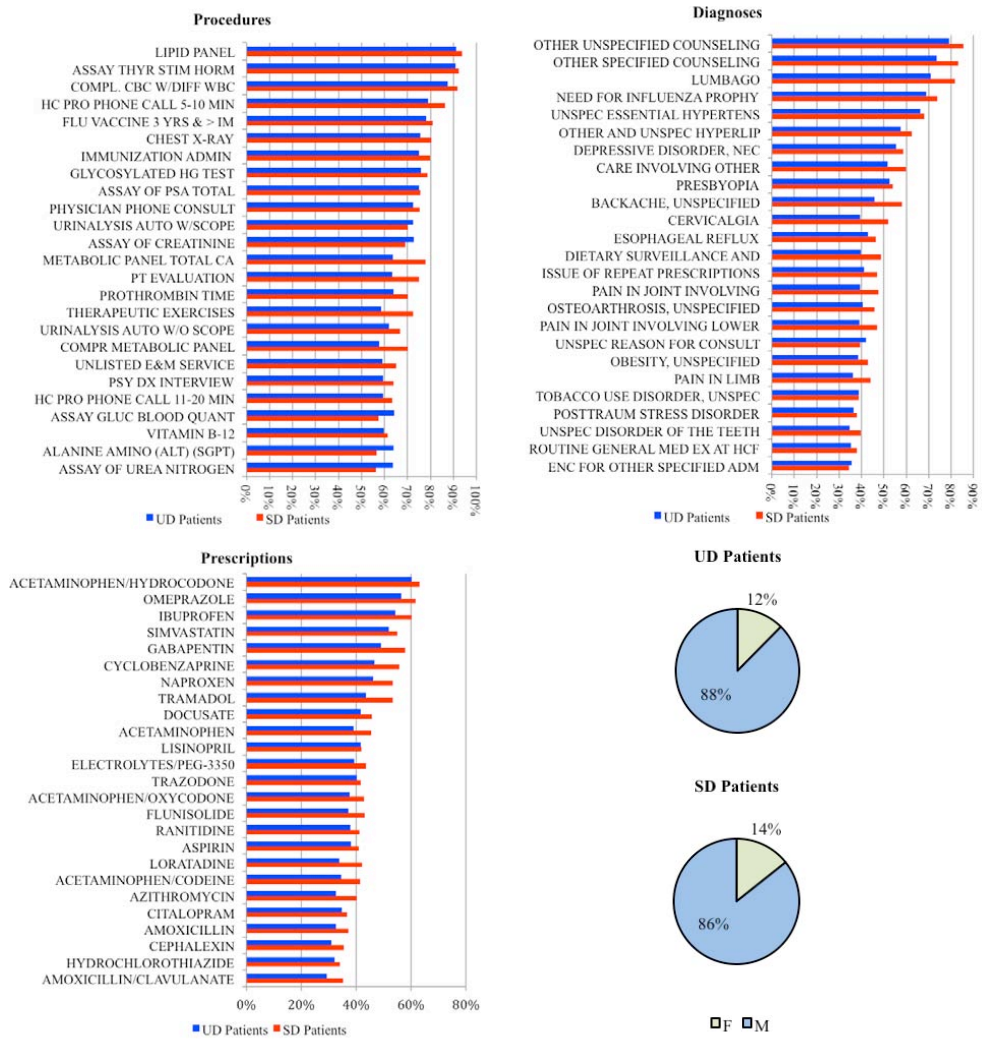

Figure 5. Percent of UD and SD patients with the most frequent procedures (by CPT code), diagnoses (by ICD9 code), and prescriptions, and gender distribution of UD and SD patients.

Table 2. Average per-patient procedure, diagnosis, prescription, and outpatient visit rates for UD and SD patients.

	Procedures per Patient	Diagnoses per Patient	Prescriptions per Patient	Visits per Patient
UD Patients	634	473	202	483
SD Patients	724	568	232	562

Discussion

In this study we identified patients in the VHA system being treated with acupuncture. We identified the cohorts using structured data and unstructured full-text data. We used CPT codes to identify patients for the structured data cohort, and SVM classification of unstructured free-text clinical notes to identify patients in the unstructured data cohort. There was a large overlap in the two sets, with only 13% of structured data patients not also being present in the unstructured data set. However, 72% of the unstructured data patients were not present in the structured data set, demonstrating the ability to significantly enlarge the set of identified acupuncture patients by using unstructured data. Our study shows that while it is feasible to identify acupuncture cohorts through structured and unstructured data independently, combining the two approaches can maximize the cohort size. This finding is consistent with findings reported by prior studies (8-11), but we show a more dramatic increase due to this medical domain not being traditionally included in electronic health records.

Aside from increasing the cohort size, combining structured and unstructured data can lead to a more representative patient population. Some prior studies compared sensitivity and specificity of different cohort identification methods, while we compared the cohorts. In comparing the cohort characteristics, we found a high degree of similarity but also some meaningful differences. Geographically, large acupuncture patient populations tend to locate in or near large metropolitan centers. The unstructured data cohort had a much higher proportion in the northwest region of Oregon, and those in the structured data cohort were proportionally more highly represented in the New York City metropolitan area. Some large metropolitan centers showed low acupuncture populations from either method. This suggests a variance in practice and/or documentation, although there are many other possibilities that will require further study to identify.

The distribution of ages in the two cohorts showed no significant difference, with the mean patient age at the time of treatment being about 56 years old in both groups. The gender representation in the two cohorts was also very similar. The rankings of the most frequent medical procedures, diagnoses, and prescriptions were very similar between the two cohorts, however there were consistently higher percentages of patients in the structured data cohort that received each procedure, diagnosis, and prescription. A frequent application of acupuncture treatment is for pain management (12), which is reflected in the frequent use of pain management prescriptions and diagnoses related to pain conditions in both cohorts.

Our data suggest that the patients in the structured data cohort had consistently higher rates of procedures, diagnoses, and prescriptions in general, not only in the most frequent sets. Patients in the structured data set also had a higher average outpatient visit rate. This indicates a difference in medical resource usage pattern between the two cohorts, with those in the structured data cohort consistently displaying higher use of VHA resources. This may indicate that patients identified only through unstructured data are relatively healthy or relying less on VHA as the sole provider.

Acknowledgements

This work is funded by VA grants CHIR HIR 08-374 and VINCI HIR-08-204.

References

1. Witt CM, Jena S, Selim D, Brinkhaus B, Reinhold T, Wruck K, et al. Pragmatic randomized trial evaluating the clinical and economic effectiveness of acupuncture for chronic low back pain. American journal of epidemiology. 2006;164(5):487-96.
2. Linde K, Vested Madsen M, Gøtzsche PC, Hrobjartsson A. Acupuncture Treatment for Pain: Systematic Review of Randomised Clinical Trials with Acupuncture, Placebo Acupuncture, and no Acupuncture Groups. Deutsche Zeitschrift für Akupunktur. 2010;53(2):40-1.
3. Yuan J, Purepong N, Kerr DP, Park J, Bradbury I, McDonough S. Effectiveness of acupuncture for low back pain: a systematic review. Spine. 2008;33(23):E887-E900.
4. Jacobson BC, Gerson LB. The inaccuracy of ICD-9-CM Code 530.2 for identifying patients with Barrett's esophagus. Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus / ISDE. 2008;21(5):452-6.

5. Lin J, Jiao T, Biskupiak JE, McAdam-Marx C. Application of electronic medical record data for health outcomes research: a review of recent literature. Expert review of pharmacoeconomics & outcomes research. 2013;13(2):191-200.
6. Friedlin J, McDonald CJ. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2006:269-73.
7. Friedlin J, Grannis S, Overhage JM. Using natural language processing to improve accuracy of automated notifiable disease reporting. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008:207-11.
8. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis care & research. 2010;62(8):1120-7.
9. Li L, Chase HS, Patel CO, Friedman C, Weng C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008:404-8.
10. Jeff Freidlin D, Marc Overhage MD P, Mohammed A Al-Haddad M, Joshua A Waters M, J. Juan R Aguilar-Saavedra M, Joe Kesterson M, et al., editors. Comparing Methods for Identifying Pancreatic Cancer Patients Using Electronic Data Sources. AMIA 2010 Symposium; 2010.
11. Elkin PL, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma H, et al. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008:172-6.
12. Vickers AJ, Cronin AM, Maschino AC, Lewith G, MacPherson H, Foster NE, et al. Acupuncture for chronic pain: individual patient data meta-analysis. Archives of internal medicine. 2012;172(19):1444-53.

CHAPTER 5

INFORMATICS CAN IDENTIFY SYSTEMIC SCLEROSIS (SSC) PATIENTS AT RISK FOR SCLERODERMA RENAL CRISIS

Reprinted from Computers in Biology and Medicine, Vol. 53, Redd D, Frech TM, Murtaugh MA, Rhiannon J, Zeng QT, Informatics can identify systemic sclerosis (SSc) patients at risk for scleroderma renal crisis, Pages 203-5, Copyright 2014, with permission from Elsevier.

Computers in Biology and Medicine 53 (2014) 203–205

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Informatics can identify systemic sclerosis (SSc) patients at risk for scleroderma renal crisis ☆

Doug Redd [a,b,c], Tracy M. Frech [a,b,*], Maureen A. Murtaugh [a,b], Julia Rhiannon [d], Qing T. Zeng [a,b,c]

[a] Veterans Affair Medical Center Salt Lake City Health Care System, Salt Lake City, Utah, USA
[b] Department of Internal Medicine, Division of Rheumatology, University of Utah School of Medicine and Veterans Affair Medical Center, Sat Lake City, Utah, USA
[c] Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, Utah, USA
[d] Veterans Affair Medical Center Denver Health Care System, Denver, Colorado, USA

ARTICLE INFO

Article history:
Received 21 April 2014
Accepted 29 July 2014

Keywords:
Informatics
Systemic sclerosis
Scleroderma
Renal crisis
Prednisone
Steroid
Blood pressure
Hypertension
Natural language processing
Management

ABSTRACT

Background: Electronic medical records (EMR) provide an ideal opportunity for the detection, diagnosis, and management of systemic sclerosis (SSc) patients within the Veterans Health Administration (VHA). The *objective* of this project was to use informatics to identify potential SSc patients in the VHA that were on prednisone, in order to inform an outreach project to prevent scleroderma renal crisis (SRC).
Methods: The electronic medical data for this study came from Veterans Informatics and Computing Infrastructure (VINCI). For natural language processing (NLP) analysis, a set of retrieval criteria was developed for documents expected to have a high correlation to SSc. The two annotators reviewed the ratings to assemble a single adjudicated set of ratings, from which a support vector machine (SVM) based document classifier was trained. Any patient having at least one document positively classified for SSc was considered positive for SSc and the use of prednisone $\geq$ 10 mg in the clinical document was reviewed to determine whether it was an active medication on the prescription list.
Results: In the VHA, there were 4272 patients that have a diagnosis of SSc determined by the presence of an ICD-9 code. From these patients, 1118 patients (21%) had the use of prednisone $\geq$ 10 mg. Of these patients, 26 had a concurrent diagnosis of hypertension, thus these patients should not be on prednisone. By the use of natural language processing (NLP) an additional 16,522 patients were identified as possible SSc, highlighting that cases of SSc in the VHA may exist that are unidentified by ICD-9. A 10-fold cross validation of the classifier resulted in a precision (positive predictive value) of 0.814, recall (sensitivity) of 0.973, and f-measure of 0.873.
Conclusions: Our study demonstrated that current clinical practice in the VHA includes the potentially dangerous use of prednisone for veterans with SSc. This present study also suggests there may be many undetected cases of SSc and NLP can successfully identify these patients.

1. Introduction

Systemic sclerosis (SSc; scleroderma) is a rare, complex autoimmune disease in which a poor prognosis is most closely related to organ fibrosis and/or hypertensive crisis. Hypertension is the key factor in the development of sudden kidney failure in SSc, which is called scleroderma renal crisis (SRC) [1]. SRC is characterized by malignant hypertension and oliguric/anuric acute renal failure, and occurs in 2% to 5% of patients with systemic sclerosis (SSc) [2]. If SRC occurs there is a 5-year survival rate of 65%, thus this condition is important to identify and prevent in SSc patients.

Several retrospective studies have found that a significant, but perhaps not widely recognized risk factor for SRC is recent use of prednisone [2,3]. The use of prednisone for any indication in SSc remains controversial, but should only be used at the lowest possible dose (ideally < 10 mg) and reserved for myositis, arthritis, interstitial lung disease, and inflammatory skin disease [4]. SSc patients should be

☆Funding: Dr. Frech was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant 8UL1TR000105 (formerly UL1RR025764). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.
* Corresponding author at: Department of Internal Medicine, Division of Rheumatology, University of Utah School of Medicine, 4B200 SOM 30 N 1900 E, Salt Lake City, 84108 Utah, USA. Tel.: +801 581 4334; fax: +801 581 6069.
E-mail address: tracy.frech@hsc.utah.edu (T.M. Frech).

http://dx.doi.org/10.1016/j.compbiomed.2014.07.022
0010-4825/

educated to monitor their blood pressure and to take the new onset of hypertension seriously [5]. It is critical to detect SRC in its earliest stages because prompt treatment with the BP-lowering class of drugs called angiotensin-converting enzyme inhibitors (ACE-inhibitors) can help to prevent progression to serious kidney failure [6].

Unfortunately, over half of cases of SRC have a delay in diagnosis, require dialysis and long-term mortality remains significant [7]. Use of prednisone in a SSc patient and/or a delay in diagnosis of SRC unfortunately can result in high morbidity and mortality due to unnecessary delays in the referral process [8].

Electronic medical records (EMR) provide an ideal opportunity for the detection, diagnosis, and management of SSc patients within the Veterans Health Administration (VHA). The VHA has one of the largest integrated healthcare systems in the United States with 8.6 million total enrollees in 2012. The VA Health Services Research and Development (HSR&D) office funded the Veterans Informatics and Computing Infrastructure (VINCI), which began operations in June 2008. VINCI is collaboration between the Office of Information and Technology (OI&T) and the Office of Research and Development (OR&D). VINCI was created to serve the data and Information Technology (IT) needs of the VA research community. The VINCI database is an excellent example of big data, a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. At the time of this study, VINCI provides access to structured and unstructured electronic medical data on 17,543,172 unique patients. VINCI and Consortium for Healthcare Informatics Research (CHIR) have shown that the EMR can be effectively utilized in a de-identified manner for patient safety and quality measurement [9].

The objective of this project was to use informatics to identify SSc patients in the Veterans Health Administration (VHA) that were prescribed prednisone in order to inform an outreach project to prevent SRC. The investigators had two goals with this informatics project: (1) to identify current SSc patients that may be inappropriate therapy, and (2) to identify if there are potential SSc patients that are not identified by ICD-9 code in order to better understand the potential impact of an outreach project.

2. Methods

The electronic medical data for this study came from VINCI and was approved for use by the Institutional Review Board. No human subjects were contacted during this research. The data consisted of structured (i.e. problem lists, medication lists, lab reports, demographics) as well as unstructured (i.e. clinical notes) data. For natural language processing (NLP) analysis, a set of retrieval criteria was developed for documents expected to have a high correlation to SSc. These criteria were: patient had at least one systemic sclerosis diagnosis by a rheumatologist or at least two diagnoses by a primary care provider and the document written by that provider contained the text “systemic sclerosis” or “scleroderma.” Snippets containing “systemic sclerosis” or “scleroderma” were extracted from these documents, manually reviewed by two annotators, and assigned ratings of “Yes”, “No”, or “Uncertain” for the positive indication of SSc. In rating the snippets, additional terms that are strong indicators of SSc based on the classification criteria for this condition [10]: “skin thickening of fingers”, “digital tip ulcers”, “fingertip pitting scars”, “telangiectasia”, “abnormal nailfold capillaries”, “pulmonary arterial hypertension”, “interstitial lung disease”, “Raynaud's phenomenon”, “anticentromere”, “anti-topoisomerase”, “anti-RNA polymerase III”, or “scleroderma-related autoantibodies” were used. The antinuclear antibody (ANA) status was not reviewed.

The two annotators reviewed the ratings to assemble a single adjudicated set of ratings, from which a support vector machine (SVM) based document classifier was trained. Any patient having at least one document positively classified for SSc was considered positive for SSc. Once a patient was confirmed as definite SSc, the use of prednisone $\geq$ 10 mg in the clinical document was manually reviewed to determine whether it was an active medication on the prescription list.

3. Results

In the VHA, there were 4272 patients that have a diagnosis of SSc determined by the presence of an ICD-9 code; all of these patients records were available for review. From a search of these patients, 1118 patients (21%) had the use of prednisone $\geq$ 10 mg documented in the EMR. Of these 1118 SSc patients on steroid, 26 had a concurrent diagnosis of hypertension and no clear plan educating the patient to monitor their blood pressure confirmed by manual review. Thus, these 26 patients were potentially being managed inappropriately.

From this manual review, the average age of the sample was 63 years. In this population 63% were confirmed to be on prednisone. Three patients were prescribed this therapy for gout; all others were prescribed this medication “for SSc” per the medical record. Only in 11 cases was indication specified as lung disease, tenosynovitis, or arthritis. No cases of myositis were identified. Prednisone doses as high as 60 mg were recorded in the EMR. The indication for the use of high rather than low dose prednisone was not clear. In the medical records of 37% of the SSc patients that were not on prednisone, phrases such as “allergy to prednisone,” “localized scleroderma,” “patient advised not to use prednisone,” or “past use of prednisone” were documented. Document types including anesthesia notes and disability forms were universally not accurate due to negation terms used, such as “this patient does not have scleroderma.”

There are limitations to our approach. We did not specifically look at ANA status because it is not a part of systemic sclerosis classification criteria, however this would have been helpful for understanding potential SSc cases. For the manual review, if scleroderma specific antibody information did not appear in the note, we could not confirm this data. Thus, low dose prednisone use in an anti-centromeric antibody SSc patient with long standing disease may not be a dangerous practice pattern, but is not captured by our study. Additionally, the study design does not adequately capture the prevalence of hypertension in this SSc population since we did not adjust for age, gender, and race/ethnicity. Patients that had both SSc (ICD-9 710.1) and localized scleroderma patients (ICD-9 701.0) coded were not excluded until NLP was applied. In this stage, only 3 patients were identified as having localized scleroderma. Thus this present study which suggests there may be many undetected cases of SSc and NLP can successfully identify these patients does not adequately exclude localized SSc. However, we demonstrated that NLP does have the ability to distinguish between these two diseases by negation terminology.

By the use of NLP an additional 16,522 patients were identified as possible SSc from all VHA records all multiple centers. From these 16,522 patients, the two annotators reviewed and rated 244 snippets. Snippets were selected by chronological date and were from multiple VAMC throughout the United States. A 10-fold cross validation of the classifier resulted in a precision (positive predictive value) of 0.814, recall (sensitivity) of 0.973, and f-measure of 0.873. Using this classifier the entire set of documents meeting the criteria was classified.

Conclusions: Current rheumatology guidelines emphasize early detection and effective management of SRC and highlight the risk of prednisone for SSc patients [3,11]. Our study demonstrated that current clinical practice in the VHA includes the potentially dangerous use of prednisone for veterans with SSc, including the

use of high dose steroid as well as unclear indications for its use. Medical plans from providers managing SSc patients with hypertension and on prednisone, did not document that patients were informed to track their blood pressure. Our study also identified many additional patients by NLP that may have a diagnosis of SSc, but were not identified by ICD-9 coding. While the absence of serology, skin score information, and patient reported outcomes in the standard note structure is a limitation to this study, the snippets of diagnostic information, did allow for identification of SSc patients. This study suggests that there may be an unusually high prevalence of SSc among veterans, which warrants further investigation.

Advances in informatics allow identification of SSc patients potentially at risk for SRC and provides the opportunity to improve quality of care in these patients through education to clinical providers. This present study also suggests there may be many undetected cases of SSc and NLP can successfully identify these patients. Manual review of cases can help providers restrict the search terms to train a classifier which will recognize phrases, such as “patient advised not to use prednisone”, in order to implement alerts appropriately. NLP may allow better identification of possible SSc patients and aid providers in the care of US veterans.

The authors have no conflicts of interests.

Conflicts of interest

All of our authors Doug Redd, Tracy M. Frech, Maureen A. Murtaugh, Julia Rhiannon, and Qing T. Zeng have declared no conflicts of interest.

References

[1] C.P. Denton, G. Lapadula, L. Mouthon, U. Muller-Ladner, Renal complications and scleroderma renal crisis, Rheumatology (Oxford) 48 (Suppl 3) (2009) Siii32–Siii35.
[2] G. Bussone, A. Berezne, V. Pestre, L. Guillevin, L. Mouthon, The scleroderma kidney: progress in risk factors, therapy, and prevention, Curr. Rheumatol. Rep. 13 (2011) 37–43.
[3] L. Guillevin, A. Berezne, R. Seror, et al., Scleroderma renal crisis: a retrospective multicentre study on 91 patients and 427 controls, Rheumatology (Oxford) 51 (2012) 460–467.
[4] D. Perez Campos, M. Estevez Del Toro, A. Pena Casanovas, P.P. Gonzalez Rojas, L. Morales Sanchez, A.R. Gutierrez Rojas, Are high doses of prednisone necessary for treatment of interstitial lung disease in systemic sclerosis? Reumatol. Clin. 8 (2012) 58–62.
[5] T.M. Frech, J. Penrod, M.J. Battistone, A.D. Sawitzke, B.M. Stults, The prevalence and clinical correlates of an auscultatory gap in systemic sclerosis patients, Int. J. Rheumatol. 2012 (2012) 590845.
[6] V.K. Shanmugam, V.D. Steen, Renal disease in scleroderma: an update on evaluation, risk stratification, pathogenesis and management, Curr. Opin. Rheumatol. 24 (2012) 669–676.
[7] H. Penn, C.P. Denton, Diagnosis, management and prevention of scleroderma renal disease, Curr. Opin. Rheumatol. 20 (2008) 692–696.
[8] H.W. Farber, R.W. Simms, R. Lafyatis, Care of patients with scleroderma in the intensive care setting, J. Intens. Care Med. 25 (2010) 247–258.
[9] J. Mayer, S. Shen, B.R. South, et al., Inductive creation of an annotation schema and a reference standard for de-identification of VA electronic clinical notes, AMIA Annu. Symp. Proc./AMIA Symp. AMIA Symp. 2009 (2009) 416–420.
[10] F. van den Hoogen, D. Khanna, J. Fransen, et al., classification criteria for systemic sclerosis: an American college of rheumatology/European league against rheumatism collaborative initiative, Ann. Rheum. Dis. 2013 (72) (2013) 1747–1755.
[11] G. Trang, R. Steele, M. Baron, M. Hudson, Corticosteroids and the risk of scleroderma renal crisis: a systematic review, Rheumatol. Int. 32 (2012) 645–653.

CHAPTER 6

AUTOMATED LEARNING OF TEMPORAL EXPRESSIONS

Reprinted from Studies in Health Technology and Informatics, Volume 216: MEDINFO 2015: eHealth-enabled Health. Douglas Redd, YiJun Shao, Jing Yang, Guy Divita, Qing Zeng-Treitler. Automated learning of temporal expressions. Pages 639-642.

 doi:10.3233/978-1-61499-564-7-639.

MEDINFO 2015: eHealth-enabled Health 639
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-639

Automated Learning of Temporal Expressions

Douglas Redd[a,b] MS, YiJun Shao[a,b] PhD, JingYang [a,b] BS, Guy Divita [a,b] MS, Qing Zeng-Treitler [a,b] PhD

[a] Salt Lake City Veteran Affairs Hospital, Salt Lake City, Utah, USA, [b] University of Utah, Salt Lake City, Utah, USA

Abstract

Clinical notes contain important temporal information that are critical for making clinical diagnosis and treatment as well as for retrospective analyses. Manually created regular expressions are commonly used for the extraction of temporal information; however, this can be a time consuming and brittle approach. We describe a novel algorithm for automatic learning of regular expressions in recognizing temporal expressions.

Five classes of temporal expressions are identified. Keywords specific to those classes are used to retrieve snippets of text representing the same keywords in context. Those snippets are used for Regular Expression Discovery Extraction (REDEx). These learned regular expressions are then evaluated using 10-fold cross validation. Precision and recall are very high, above 0.95 for most classes.

Keywords:

Electronic Medical Record, Machine Learning

Introduction

Temporal expressions are a special type of named entity. In clinical notes, temporal information is often critical to the interpretation of findings. The time order of events is an essential factor in assessing causation and in evaluating co-occurrence [1]. For example, an adverse reaction to a treatment can only be shown if the reaction occurred after the treatment. Additionally, drug-drug interactions can only be shown if the drugs were taken within the same time frame.. In prior studies that extracted temporal expressions, regular expressions are commonly used [2]. While regular expressions are powerful, they do typically need to be manually created based on chart reviews. This creates a challenge for maintenance and adaptation. In this paper, we describe the use of a simple and novel learning algorithm to discover temporal expressions. While the expressions we discovered are specific to the set of training data employed by this study, the algorithm is generalizable to other training datasets and to other types of expressions.

Background

The temporal information in medical records is important for both clinical decision support and general medical research. This information exists as both structured data and unstructured narrative data. Extraction of temporal information from structured data is trivial but is much harder from the narrative data [3]. For instance, temporal expression extraction was part of the task proposed by the 2012 i2b2 NLP for Clinical Data challenge [4].

Earlier studies have attempted to identify the temporal information using NLP techniques in a simplified form, e.g., “historical” vs. “current” [5,6]. More recent studies have shown that it is feasible to extract the exact temporal expressions from clinical narratives. For example, in one study [7] a system called Med-TTK was developed for detecting temporal expressions in medical narratives by extending an existing system called Temporal Awareness and Reasoning Systems for Question Interpretation Toolkit (TTK) [8]. TTK is an open-source software package developed for extracting temporal information in news articles. The Med-TTK system achieved an overall F-score of 0.85 on a set of 200 U.S. Department of Veterans Affairs (VA) clinical notes. Another study [9] combined rules and machine learning for the extraction of temporal expressions and achieved a micro F-score of 0.90 on a set of 320 clinical notes provided by the 2012 i2b2 challenge.

Methods

Corpus

We used a Pittsburgh EHR dataset (Pitts Data) for this study [10]. Pitts Data is composed of 100,866 de-identified clinical notes and was used in the Text Retrieval Conference (TREC) Medical Track [11], an annual event for encouraging the development of medical information retrieval techniques.

Keyword Identification

An initial set of temporal keywords were constructed from TIMEN, a temporal expression normalization tool [12]. We used 193 keywords that included time of a day, month, season, decade, and holidays. Sample keywords are shown in table 1.

Table 1 – Sample Temporal Expression Keywords

at	pm
after	midday
before	afternoon
between	noon
by	evening
during	overnight
following	midnight
for	night
from	pm
on	p m
since	pm
till	Seconds
to	second
until	second
within	minutes
while	minute
when	Hours
except	hour
in	days
a.m.	day

Snippet Extraction

The set of temporal keywords were used to extract a set of snippets from the document corpus. We used snippets to refer to the keyword and the context surrounding the keyword in the document. In this study, each snippet consisted of the matched keyword, the 10 words preceding the keyword in the document, and the 10 words following the keyword in the document. We progressively extracted all snippets containing the keywords from documents until we surpassed our target of 1,000 snippets for annotation. This resulted in an annotation set of 1,008 snippets from 13 documents.

Annotation

Human annotation using VTT [13] was performed to indicate the exact temporal expression in each snippet and the class of the temporal expressions (figure 1). The classes were DATE, TIME, DURATION, SET, and OTHER, with examples given in table 2. The temporal expression was marked in each snippet (referred to as the "labeled segment"), and it's temporal class assigned, by two independent annotators. To reach agreement, two rounds of annotation testing were performed on subsets of data. The kappa for the second round was 94%. Afterwards one annotator performed the annotation for the remaining documents.

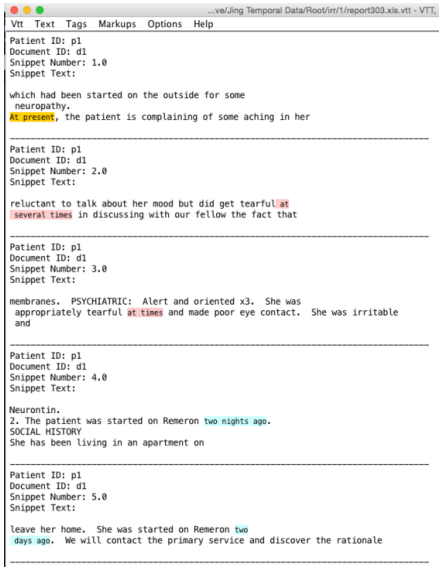

Figure 1 – Sample of VTT annotations.

Information Extraction

We trained a regular expression discovery Extraction (REDEx) algorithm on the annotated snippets. The REDEx algorithm is a novel process we have developed that automatically learns regular expressions that capture the value of the annotated labeled segment and the context surrounding it. We have used previous versions of REDEx for various value extraction tasks, including body weight [14]. We implemented REDEx in Java, making extensive use of the java.util.regex.* core libraries. It is open source and licensed under the Apache License, Version 2.

Table 2 – Temporal Expression Classes

Class	Description	Examples
DATE	Regarding a specific day.	Jan 14 2007
TIME	A specific time point.	14:04:28
DURATION	A period of time.	40 minutes
SET	A set of several temporal expressions.	Monday and Wednesday
OTHERS	Temporal expressions with vague resolution.	past

The current version of REDEx is represented as pseudo-code in Figure 3. Briefly, each annotated snippet was first split into 3 parts: the "labeled segment (LS)", which is the piece of text marked by the annotator; the "before labeled segment (BLS)", which is the text between the beginning of the snippet and the start of the LS; and "after labeled segment (ALS)", consisting of the text between the end of the LS and the end of the snippet (figure 2). Each BLS-LS-ALS triplet was then generalized to a regular expression by replacing all punctuation, whitespace, and digits with corresponding regular expressions: \p{Punct}, \s{1,50}, and \d+ respectively. The BLS-LS-ALS triplets were then iteratively generalized by successive rounds of trimming from the front of the BLS and the end of the ALS until one or more false positives were observed. We interpreted a false positive to be a case were REDEx predicted a value where there was none in the manual annotations. Duplicate triplets were removed, and then the triplets were combined into a single regular expression. The LS was marked as a capture group in order to retrieve the matched value. Sensitivity was then calculated for each regular expression by dividing the count of matched snippets by the total number of snippets.

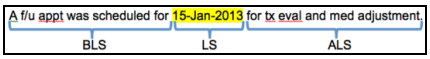

Figure 2 - Before Labeled Segment (BLS), Labeled Segment (LS), and After Labeled Segment (ALS) of a date expression in a phrase.

Evaluation

Evaluation of the regular expressions was performed using 10-fold cross validation, with final scores being the mean of those from each of the folds. Evaluation measures of precision, recall, F1-score, and accuracy were calculated for each class, along with the number of snippets in each class and the number of unique regular expressions generated by REDEx. We defined predictions as true positive (TP) when the regular expressions extracted a value from a snippet that matched the annotated value, false positive (FP) when the extracted value did not match the annotated value, false negative (FN) if the regular expressions failed to extract a value but there was an annotated value, and true negative (TN) if there was not an extracted value and there was no annotated value.

Results

1,008 snippets were extracted for annotation using 193 keywords from the corpus. These snippets were human annotated to identify the temporal expression spans and classified as DATE, TIME, DURATION, SET, or OTHER. Regular expressions and their sensitivities were then automatically generated using the RED extraction algorithm for each class. Samples of the resulting regular expressions are shown in table 3.

Evaluation measures of precision, recall, F1-score, accuracy, number of snippets in each class, and the number of unique regular expressions generated by REDEx are presented in table 4. Evaluation metrics were very high in most cases. Date, Time, and Set classes were all ≥ 0.97 for precision. Recall was ≥ 0.96 for Date, Time, Duration, and Other. The only measures below 0.90 were for the Set class, where recall and accuracy were 0.83 and the sample size was very small at 18.

Table 3 – Examples of REDEx Regular Expressions

Class	Regular Expression
DATE	(ten\s{1,50}days\s{1,50}ago) \p{Punct}DATE\p{Punct}(Aug\s{1,50}\d+\s{1,50}\d+)
TIME	(\d+\p{Punct}\d+\p{Punct}\d+\s{1,50}AM) \s{1,50}\d+\s{1,50}\d+\p{Punct}\s{1,50}(\d+\p{Punct}\d+\p{Punct}\d+)\s{1,50}T
DURATION	\s{1,50}(another\s{1,50}four\s{1,50}weeks) \s{1,50}\S{1,6}\s{1,50}\S{1,6}\s{1,50}\S{1,10}\p{Punct}\s{1,50}on\s{1,50}a\s{1,50}\S{1,5}\s{1,50}\S{1,5}\s{1,50}\S{1,5}\p{Punct}\s{1,50}over\s{1,50}(the\s{1,50}past\s{1,50}few\s{1,50}months)
SET	(every\s{1,50}\d+\s{1,50}hours) (per\s{1,50}hour)
OTHERS	(recently)\s{1,50}discharged\s{1,50}from\s{1,50}the\s{1,50}hospital\s{1,50}and\s{1,50}\S{1,10}\s{1,50} \s{1,50}\S{1,4}\s{1,50}minimally\s{1,50}invasive\s{1,50}esophagectomy\s{1,50}with\s{1,50}reanastomosis\s{1,50}known\s{1,50}to\s{1,50}our\s{1,50}service\s{1,50}for\s{1,50}(recent)

Table 4 – Evaluation Metrics

Measure	Date	Time	Duration	Set	Other
Precision	0.97	0.98	0.93	1.00	0.93
Recall	0.97	0.97	0.96	0.83	0.96
F1-score	0.97	0.98	0.95	0.91	0.95
Accuracy	0.94	0.95	0.90	0.83	0.90
# Snippets	493	289	169	18	39
# Reg Ex	128	54	41	7	17

```
BLS = Before Labeled Segment
LS = Labeled Segment
ALS = After Labeled Segment

PS = Positive Text Snippet
NS = Negative Text Snippet

Regular Expression Discovery (PS, NS)

/*Initialize Result*/
RS = { }

/*For each positive instances*/
For each p in PS
        /*Transform a text string into regular expression
        by replacing punctuations, white spaces and dig-
        its*/
        P' = Generalize (p);
        /*Split each expression into 3 segments*/
        (bls', ls', als') = Split (p');
End For

/*Combine all labeled patterns*/
ls_exp = Combine (LS');

/*For each positive instance*/
For each p
        /*Test the regular expression using the negative
        instances; if an expression matches any negative
        instances, it is discarded. */
        While match (p', NS) == False
                /*Trim the segments before or after the
                labeled segments*/
                p'' = trim (bls', ls', als');
        End While

        /*Add the shortest regular expression that does
        NOT match any snippets in the negative sam-
        ple*/
        RS = RS + p';
End For
```

Figure 3 - Pseudo-code describing the RED Extraction algorithm

Discussion

This study has demonstrated the feasibility of automatically discovering temporal expressions. Given the moderate amount of training data, we were able to achieve a very high level of sensitivity and specificity. The machine-generated expressions are humanly readable, though often not as succinct as the expression a senior programmer would have written. It is also worth noting that, the REDEx algorithm does not require seed patterns to begin with.

The REDEx algorithm can be applied to other use cases as well. We have, for instance, used REDEx to extract weight value and unit, with the same level of sensitivity and specificity as for temporal extraction [14]. When using REDEx, we trade the time required for writing and testing regular expression with the time required for annotation. In the use cases of temporal information and weight value/unit, we found the trade-off to be a beneficial. In these two cases,

research assistants and researchers who do not have programming knowledge did the annotations very quickly (100 to 150 annotations/hour after the inter rater agreement is established). The amount of time it takes REDEx to generate expressions was trivial comparing to the manual generation of expressions, thus REDEx can potentially incorporate many more examples. The development of REDEx itself did take time. It, however, can be re-used. For the purpose of future maintenance, we felt that it would be much easier to create additional annotations than to manually revise the expressions.

There are many machine learning algorithms for text classification, including for context classification. There are considerably fewer learning algorithms that we can readily use for the discovery of specific (regular) patterns. With REDEx, we can learn to recognize sequential patterns with numbers, symbols and letters. Although developed and evaluated using the English alphabet, this should be directly applicable to other Latin alphabets using Arabic numerals.

The Pitts dataset used for training and testing is relatively uniform. Our concern is that it does not contain sufficient variations of the temporal expressions. We plan to extend the learning and testing to other clinical notes and create a larger set of temporal expressions. In addition, we plan to evaluate the REDEx algorithm in comparison to manually created regular expressions and possibly the Med-TTK. Other future experiments are to evaluate with other corpora, and to determine the accuracy of regular expressions discovered in one corpus when applied to a different corpus.

We performed the progressive snippet selection method which did not differentiate between the temporal classes at the time of selection. This resulted in small sample sizes for the Set and Other classes. In future studies the selection method can be improved to provide better representation of all classes. This can be done using a similar progressive selection method, but counting the snippets separately for each class. Or by extracting all snippets, stratifying them by class, then randomly sampling an equal number from each stratum. This would have the added advantage of providing class cardinality across the corpus as well as better representation of samples.

The temporal expressions discovered in this study have been incorporated into a v3NLP Framework NLP module [15] to look up temporal text from clinical notes. Additional classes of temporal expressions, in particular intervals, are planned as additions. When extended and tested on other datasets, the NLP module will be released as open source software.

Conclusion

We have developed and tested a simple and novel REDEx algorithm for the discovery of temporal expressions, with good sensitivity and specificity. The REDEx can be applied to other types of information extraction tasks, because it does not contain any built-in information about temporal data.

Acknowledgements

This work was funded by VA grants: NLP Ecosystem CRE 12-315, CHIR HIR 08-374 and VINCI HIR-08-204.

References

[1] Hill AB. The Environment and Disease: Association or Causation? Proc R Soc Med. 1965 May;58(5):295-300. PubMed PMID: 14283879.

[2] Babbar R, Singh N. Clustering based approach to learning regular expressions over large alphabet for noisy unstructured text. Proceedings of the fourth workshop on Analytics for noisy unstructured text data; Toronto, ON, Canada. 1871848: ACM; 2010. p. 43-50.

[3] Zhou L, Hripcsak G. Temporal reasoning with medical data—A review with emphasis on medical natural language processing. Journal of Biomedical Informatics. 2007 4//;40(2):183-202.

[4] Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. Journal of the American Medical Informatics Association. 2013;20(5):806-13.

[5] Chapman WW, Chu D, Dowling JN. ConText: an algorithm for identifying contextual features from clinical text. Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing; Prague, Czech Republic. 1572408: Association for Computational Linguistics; 2007. p. 81-8.

[6] Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics. 2009 10//;42(5):839-51.

[7] Reeves RM, Ong FR, Matheny ME, Denny JC, Aronsky D, Gobbel GT, et al. Detecting temporal expressions in medical narratives. International Journal of Medical Informatics. 2013 2//;82(2):118-27.

[8] Verhagen M, Mani I, Sauri R, Knippen R, Jang SB, Littman J, et al. Automating temporal annotation with TARSQI. Proceedings of the ACL 2005 on Interactive poster and demonstration sessions; Ann Arbor, Michigan. 1225774: Association for Computational Linguistics; 2005. p. 81-4.

[9] Kovačević A, Dehghan A, Filannino M, Keane JA, Nenadic G. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. Journal of the American Medical Informatics Association. 2013;20(5):859-66.

[10] Voorhees EM, Hersh W. Overview of the TREC 2012 Medical Records Track. Available from: http://trec.nist.gov/pubs/trec21/papers/MED12OVERVIEW.pdf.

[11] Edinger T, Cohen AM, Bedrick S, Ambert K, Hersh W. Barriers to Retrieving Patient Information from Electronic Health Record Data: Failure Analysis from the TREC Medical Records Track. AMIA Annual Symposium Proceedings. 2012 11/03;2012:180-8. PubMed PMID: PMC3540501.

[12] Llorens H, Derczynski L, Gaizauskas R, Saquete E, editors. TIMEN: An Open Temporal Expression Normalisation Resource 2012: European Language Resources Association (ELRA).

[13] Visual Tagging Tool: United States National Library of Medicine; 2010. VTT]. Available from: http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/vtt/current/web/index.html.

[14] Murtaugh MA, Gibson BS, Redd D, Zeng-Treitler Q. Regular expression-based learning to extract bodyweight values from clinical notes. Journal of Biomedical Informatics. (0).

[15] Divita G, Zeng-Treitler Q, Gundlapalli AV, Duvall S, Nebeker J, Samore MH. Sophia: A Expedient UMLS Concept Extraction Annotator. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2014:467-76.

Address for correspondence

Qing Zeng-Treitler <q.t.zeng@utah.edu>

CHAPTER 7

DISCUSSION

7.1 Introduction

Several different subdomains of clinical research were examined in order to investigate patient and feature identification using unstructured data. Multiple methods of information retrieval and extraction were used in various tasks, including a novel machine-learning algorithm, to demonstrate that the addition of unstructured data search to structured data search can dramatically increase cohort size as well as the number of feature observations for each patient. This data enrichment can enable studies that were not previously possible due to insufficient cohort size or number of observations.

The task of cohort identification was looked at first, and began with the process of query formation, showing it could be improved with query expansion. We addressed clinical cohort size, and demonstrated that the addition of free text search of unstructured data can increase cohort size. We also looked at differences between patient cohorts that are identified with structured and unstructured data. We continued by analyzing the clinical impact of the addition of unstructured data in the recognition of a complex and serious condition, scleroderma renal crisis. Next we examined cohort enrichment using machine learning for information extraction. Very good performance was demonstrated for a novel extraction algorithm, REDEx, in the task of extracting temporal expressions.

We went on to show meaningful increases in cohort size and number of event observations in patients using REDEx.

7.2 Cohort Identification

7.2.1 Query Formation

We first examined the task of cohort identification, and began with the process of query formation to determine if this process could be improved with query expansion. Four query expansion methods were evaluated, one of which was an ensemble of the other three. Each was evaluated using a set of 12 queries against 600 documents. Overall, we demonstrated that automated query expansion can improve results.

A large variance in precision and recall was observed among the four methods on many of the queries. This was expected due to the different methodologies used in the methods. Only two of the methods, topic model and synonyms, consistently outperformed the baseline of no expansion. An ensemble method was used as an attempt to incorporate the best qualities of the other methods into a best-of-breed style of expansion. This was mostly unsuccessful, and was likely due to the semantic distance measure we used for combining the results of the other methods. The semantic distance method rewards common terms between the methods, which likely causes an averaging rather than a maximizing effect.

By using a different method of combining results of individual methods, the ensemble method could be improved. A subsequent study using simple summing up of weights showed improved ensemble performance [1]. Additional improvement may be found by using the highest scoring terms from the different methods, regardless of their commonality. Another source of potential improvement is the use of interactive

expansion, which allows the user to choose which expansion terms to include or exclude. This has been successful in other domains, but requires user interaction [2-4]. This can be especially successful when the user has expert knowledge in the domain being searched. Relevance feedback is another approach that has been implemented successfully in other domains [5]. In this approach, after the initial retrieval, the user evaluates individual documents for their relevance to indicate that they would like to see documents "more like this." These approaches might be successfully applied to clinical note retrieval; however, they place a greater burden on the user.

7.2.2 Comparing Free Text and Structured Queries

7.2.2.1 Cohort Size

We addressed clinical cohort size, and showed that the addition of free text search of unstructured data increases cohort size. It is reasonable to expect that adding unstructured data search to structured data search will increase cohort size at least minimally. This has been demonstrated previously [6, 7]. However, the increase is of enough magnitude and accuracy to justify the additional effort required. Initial results showed dramatic increases; however, further positive predictive value analysis tempered these results somewhat. Some use cases, such as height and weight, showed small increases of under 10%. Others, such as uncontrolled diabetes showed large increases of up to 50%, while concurrent use of Ginkgo and Warfarin showed truly dramatic increases. This variance demonstrates that the magnitude of the benefit of adding free text search of unstructured data is very use case dependent. In addition to cohort size, addition of unstructured data search can increase the number of detectable observations for patients. We found that the number of height and weight observations was nearly doubled

by the addition of unstructured data. This can be important, particularly for time series studies, to show changes in measurements over time or to estimate a measurement at a particular point in time. The number and timing of observations can be important for vital signs as well as acute diagnoses.

In performing these searches, we discovered techniques that can help improve future results. False positives were more common in the searches for Ginkgo, uncontrolled diabetes, and abdominal girth. In these cases, the false positives were more highly represented in specific note types. Conversely, true positives had higher occurrence in other note types. By excluding note types known to have a high rate of false positives, performance may be improved. Alternatively, rather than excluding them completely, note types with high false positive rates could be negatively weighted, and those with high true positive rates could be positively weighted.

We also found improvements by applying natural language processing (NLP) techniques to candidate results. Detection of negations, conditionals, prescriptive phrases, and other contextual indicators may greatly improve results. This comes at the expense of a substantial increase in effort, however, and must be weighed against the magnitude of the expected benefit. In these cases, representative sampling is important to the success of the process. Very large data sets present a unique challenge in this area. Stratified sampling based on note type, geographic area, date, and other factors can improve the overall representativeness of the sample.

7.2.2.2 Cohort Characteristics

In addition to increase of cohort size, addition of patients identified through unstructured data may include patient subpopulations that are under-represented in

structured data. Acupuncture is a procedure that is not consistently recorded in electronic health records due to uncertainty regarding its effectiveness. Despite this, acupuncture has a high rate of usage and a more complete patient cohort can be important for effectiveness studies as well as the determination of usage rates to inform service capacity planning.

We found that cohorts identified by structured and unstructured data were very similar but had important differences. The geographic distribution of acupuncture users was predictably shown to have the highest density around large metropolitan areas. However, some metropolitan centers are represented predominantly in structured data, and others predominantly in unstructured data. Patients identified by structured data displayed a higher overall usage rate of resources. There are many possible explanations for this that will require further study to elucidate. Data used for this study come from the Veterans Health Administration (VHA). Patients identified by unstructured data may live in areas with lower proximity to VHA resources in general, or for VHA acupuncture services in particular. This would lead these patients to obtain acupuncture, and possibly other health services, more frequently outside of the VHA system. In these cases, acupuncture use would likely be recorded more in unstructured notes from clinicians, and a less complete overall medical record in the VHA system may result. Cultural attitudes toward acupuncture, and health services generally, may vary by geographic region. Documentation practices may vary between VHA facilities, or it may be that some geographic areas are generally healthier and require lower health service usage.

An explanation likely lies in a combination of some of these, and other, possibilities. In the VHA system, it is clear that by using a cohort exclusively derived

from structured data a subpopulation with distinct differences is excluded. A larger and more complete cohort is obtained when adding unstructured data when finding cohorts.

7.2.2.3 Clinical Impact

Examples have been documented where unstructured data improve the detection of clinical conditions [8-12]. Complex conditions may not be detectable using structured data alone. The addition of unstructured data can have a direct clinical impact by identifying risk for a complex condition such as scleroderma renal crisis (SRC). Rheumatology disorders can share many of the same symptoms, and a patient may suffer from multiple disorders concurrently. An incomplete diagnosis, where only one of many disorders is diagnosed, can result in misguided treatment. In the case of systemic sclerosis (SSc), this can lead to the dangerous condition of SRC [13]. Scleroderma renal crisis precipitated by steroid treatment in systemic lupus erythematosus and scleroderma overlap syndrome. Prednisone is often prescribed for the treatment of symptoms of rheumatology disorders. Prednisone, however, can cause SRC in patients suffering from SSc, which makes the determination of SSc important in the treatment of rheumatology disorders.

We identified the use of high dose prednisone in SSc patients in the VHA system using structured data. This problem was compounded when we used natural language processing (NLP) on unstructured data, which identified many more patients at risk of SRC than the use of structured data alone. Further systematic validation is required; however, this may indicate a need for a prescription alert system to warn prescribers of potentially harmful effects when prescribing prednisone to patients with SSc. NLP is a

necessary part of an alert system due to the incomplete recording of SSc in structured data.

7.3 Cohort Enrichment

7.3.1 Machine Learning to Extract Numerical/Categorical Entities

7.3.1.1 Machine Learning Performance

Temporal expressions in unstructured clinical data are important to clinical reasoning. The times when events and measurements occur are essential to understanding patient status. Time order of events and determination of event concurrency is required in order to evaluate causation. Previous attempts at temporal expression identification in clinical notes have suffered from problems of accuracy, adaptability, and generalizability to new data sets [14-16]. We developed a novel value extraction algorithm, regular expression discovery for extraction (REDEx), that automatically learns regular expressions from annotated documents. We applied it to the extraction of temporal expressions from clinical notes and obtained a high degree of accuracy with a relatively small training set. The regular expressions created by REDEx are human readable and usable in other regular expression systems.

Regular expressions are commonly used for information extraction; however, a human normally creates them manually. The author must have knowledge of regular expressions as well as the clinical domain being studied, which can be a rare combination in one person. Also, this creates a maintenance problem when new documents are added that may contain patterns not seen previously by the human. REDEx changes this by moving the effort to an annotation task. Domain experts annotate a corpus to indicate the values that should be extracted. The REDEx algorithm then automatically learns regular

expressions from the annotated documents. This creates a system that is more maintainable because only the domain expert is required. When new documents are added to the cohort, a subset of those can be annotated and REDEx applied in order to update the set of regular expressions.

The corpus used in this study was not large and was relatively uniform. A larger and more diverse corpus needs to be used to demonstrate generalizability and applicability to large clinical corpora. Also, the sampling method used under-represented some classes of temporal expressions. Future work will need to improve representation of all temporal classes, using stratified sampling, over/undersampling, or other methods.

7.3.1.2 Comparison of Number of Observations and Cohort Size

We analyzed the ability of REDEx to increase cohort size and find additional observations in patients already identified. To this end, we investigated bodyweight related values in clinical notes, and the effectiveness of REDEx for extraction of those values. We found a substantial increase in both a) the number of individuals with bodyweight values and b) the number of bodyweight values for each individual. These findings corroborated other studies exposing shortcomings in structured data [17-19]. The increase in unique bodyweight values in the present study is smaller than in the cited studies. The recording of patient bodyweight as structured data is mandated as one of the core objectives of meaningful use, stage 1, of the HITECH Act's EHR Incentive Program [20]. Thus, we expected that bodyweight values in structured data would have a high occurrence, and that the addition of unstructured data would have a smaller effect. The smaller effect of 7.7% is still very meaningful, especially when evaluated on very large EHR systems where the number of additional observations would be large.

A limitation is that models trained on one corpus may give unsatisfactory results on another corpus. This is a problem common to many NLP approaches. Annotated documents from the new corpus may be incorporated for adaptation of the model. REDEx can be very rigid in that no false positives are allowed during training. This can result in models that favor precision over recall. A possible solution to this is to investigate whether allowing a small number of false positives in return for more true positives can improve recall without hurting precision. An investigation of whether the addition of more observations would change the obesity classification of a significant number of individuals, or change their treatment trajectory, is also important potential work. It also would be useful to compare additional actual observations to estimate values generated from statistical imputation or other means. Estimated values may be reasonable since bodyweight values generally do not have large sudden changes over time.

7.4 Conclusion

Unstructured data in the electronic health record are a wealth of information with importance far beyond their primary use. Deficiencies in structured data have been demonstrated in many domains. Efficient analysis of unstructured data can increase the number of patients with specific characteristics that can be detected, and also give a more complete picture of the patient by detecting additional observations. Methods for optimizing the use of unstructured data remain an active research subject.

Clinical narrative is a unique challenge for text analysis. Acronyms and abbreviations, nongrammatical structures, conditionals, and negations are ubiquitous, leading to its description as a unique sublanguage.

Methods for automated query expansion were developed that greatly improve

performance of keyword-based information retrieval without additional effort required from the person performing the query. Cohorts identified in this way can still be incomplete in many areas, so we developed NLP methods for cohort identification to address this. By including unstructured data, we demonstrate that, compared to structured data alone, cohort size can be greatly increased, a more complete population can be identified, and important clinical conditions can be detected that are only minimally detectable otherwise. In addition, we found a much more complete representation of patients can be obtained. We developed a novel machine learning algorithm for information extraction, REDEx, that can efficiently extract clinical values from unstructured clinical text, adding additional information and observations over what is available in structured text alone.

A limitation of work in this domain, including this work, is assurance of completeness of the initial document retrieval. Any flaws at this stage will be compounded in later stages. The work herein dealing with query formation and query expansion improves the likelihood of obtaining a complete document set, but this is an area where attention always needs to be paid. Another potential limitation is that the REDEx algorithm is very precise; however, it can suffer when it comes to recall. To address this, I am continuing research on the REDEx algorithm to include a second tier of regular expressions targeted at recall, where if there are no matches in the high precision tier 1 regular expressions, matches against the second tier of high recall regular expressions will be attempted. Another challenge is making tools realistically available to a large community. I have addressed this with the query expansion tools we developed by making them are open source and also deploying them as publicly available web services.

My implementation of the REDEx algorithm is also available as source code, but making it available as a web service would increase its ease of use, especially to non-programmers.

There are many additional follow on studies that would be of use. The REDEx algorithm is designed to be highly adaptable for documents from new sources, which should require annotation of a relatively small set of documents from the new source. This needs to be validated with a formal study, however. Also, REDEx has only been applied to English language documents. It should be able to be successfully used for other Latin alphabet-based languages using Arabic numerals. This would greatly increase its usefulness worldwide.

The implications of this work are many. Primarily they are: the availability of larger cohorts; cohorts that are more representative of target populations; and cohorts with richer feature sets than previously achievable. This increases the capability of scientific discovery, and makes research possible that was previously not feasible.

7.5 References

1. Bui D, Redd D, Rindflesch T, Zeng-Treitler Q. An ensemble approach for expanding queries. The Twenty-Second Text REtrieval Conference (TREC 2012) Proceedings. 2012 November 6-9, 2012.

2. Kaczmarek AL. Interactive query expansion with the use of clustering-by-directions algorithm. IEEE T Ind Electron. 2011 Aug;58(8):3168-73. PubMed PMID: WOS:000293685700007.

3. Beaulieu M, Fowkes H, Alemayehu N, Sanderson M. Interactive okapi at Sheffield-TREC-8. NIST Special Publication. 2000 (246):689-6998.

4. Sahib NG, Tombros A, Ruthven I. Enabling interactive query expansion through eliciting the potential effect of expansion terms. Advances in Information Retrieval: Springer; 2010. p. 532-43.

5. Efthimiadis EN. Interactive query expansion: a user-based evaluation in a relevance feedback environment. J Am Soc Inf Sci. 2000;51(11):989-1003.

6. Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care. 2007;13(6 Part 1):281-8.

7. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008;35:128-44.

8. Garvin JH, DuVall SL, South BR, Bray BE, Bolton D, Heavirland J, et al. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. J Am Med Inform Assoc. 2012 Sep-Oct;19(5):859-66. PubMed PMID: 22437073. PMCID: 3422820.

9. Jacobson BC, Gerson LB. The inaccuracy of ICD-9-CM code 530.2 for identifying patients with Barrett's esophagus. Dis Esophagus. 2008;21(5):452-6.

10. Ding EL, Song Y, Manson JE, Pradhan AD, Buring JE, Liu S. Accuracy of administrative coding for type 2 diabetes in children, adolescents, and young adults. Diabetes Care. 2007;30(9):e98. PubMed PMID: 17726188.

11. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc. 2014 Sep-Oct;21(5):801-7. PubMed PMID: 24384230. PMCID: PMC4147606.

12. Li L, Chase HS, Patel CO, Friedman C, Weng C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. AMIA Annu Symp Proc. 2008:404-8. PubMed PMID: 18999285. PMCID: 2656007.

13. Alayoud A, Qamouss O, Hamzi AM, Benyahia M, Oualim Z. Scleroderma renal crisis precipitated by steroid treatment in systemic lupus erythematosus and scleroderma overlap syndrome. Arab J Nephrol Transplant. 2012;5(3):153-7.

14. Reeves RM, Ong FR, Matheny ME, Denny JC, Aronsky D, Gobbel GT, et al. Detecting temporal expressions in medical narratives. Int J Med Inform. 2013;82(2):118-27.

15. Verhagen M, Mani I, Sauri R, Knippen R, Jang SB, Littman J, et al. Automating temporal annotation with TARSQI. Proceedings of the ACL 2005 Interactive Poster and Demonstration Sessions. 2005;81-4.

16. Kovačević A, Dehghan A, Filannino M, Keane JA, Nenadic G. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. J Am Med Inform Assoc. 2013;20(5):859-66.

17. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res. 2005;40(5p2):1620-39.

18. Das SR, Kinsinger LS, Jr YW, Wang A, Ciesco E, Burdick M, et al. Obesity prevalence among veterans at Veterans Affairs medical facilities. Am J Prev Med. 2005;28(3):291-4.

19. Green BB, Anderson ML, Cook AJ, Catz S, Fishman PA, McClure JB, et al. Using body mass index data in the electronic health record to calculate cardiovascular risk. Am J Prev Med. 2012;42(4):342-7. PubMed PMID: PMC3308122.

20. Pagano M, Gauvreau K. Principles of biostatistics. Second edition: Duxbury; 2000.

APPENDIX

REGULAR EXPRESSION-BASED LEARNING TO EXTRACT BODYWEIGHT VALUES FROM CLINICAL NOTES

Reprinted from the Journal of Biomedical Informatics, Vol. 54, Murtaugh MA, Gibson BS, Redd D, Zeng-Treitler Q, Regular expression-based learning to extract bodyweight values from clinical notes, Pages 286-190, Copyright 2015, with permission from Elsevier.

Journal of Biomedical Informatics 54 (2015) 186–190

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Regular expression-based learning to extract bodyweight values from clinical notes

Maureen A. Murtaugh [a,b,*], Bryan Smith Gibson [a,b], Doug Redd [a,c], Qing Zeng-Treitler [a,c]

[a] IDEAS Center, Veterans Administration, Salt Lake City Health Care System, Salt Lake City, UT, United States
[b] Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, United States
[c] Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, United States

ARTICLE INFO

Article history:
Received 6 November 2014
Accepted 24 February 2015
Available online 5 March 2015

Keywords:
Natural language processing
Bodyweight
Text classification

ABSTRACT

Background: Bodyweight related measures (weight, height, BMI, abdominal circumference) are extremely important for clinical care, research and quality improvement. These and other vitals signs data are frequently missing from structured tables of electronic health records. However they are often recorded as text within clinical notes. In this project we sought to develop and validate a learning algorithm that would extract bodyweight related measures from clinical notes in the Veterans Administration (VA) Electronic Health Record to complement the structured data used in clinical research.

Methods: We developed the Regular Expression Discovery Extractor (REDEx), a supervised learning algorithm that generates regular expressions from a training set. The regular expressions generated by REDEx were then used to extract the numerical values of interest.

Methods: To train the algorithm we created a corpus of 268 outpatient primary care notes that were annotated by two annotators. This annotation served to develop the annotation process and identify terms associated with bodyweight related measures for training the supervised learning algorithm. Snippets from an additional 300 outpatient primary care notes were subsequently annotated independently by two reviewers to complete the training set. Inter-annotator agreement was calculated.

Methods: REDEx was applied to a separate test set of 3561 notes to generate a dataset of weights extracted from text. We estimated the number of unique individuals who would otherwise not have bodyweight related measures recorded in the CDW and the number of additional bodyweight related measures that would be additionally captured.

Results: REDEx's performance was: accuracy = 98.3%, precision = 98.8%, recall = 98.3%, F = 98.5%. In the dataset of weights from 3561 notes, 7.7% of notes contained bodyweight related measures that were not available as structured data. In addition 2 additional bodyweight related measures were identified per individual per year.

Conclusion: Bodyweight related measures are frequently stored as text in clinical notes. A supervised learning algorithm can be used to extract this data. Implications for clinical care, epidemiology, and quality improvement efforts are discussed.

1. Introduction

The use of Electronic Health Record (EHR) data in conjunction with data extraction and categorization tools (e.g. clinical phenotyping), holds great potential to improve clinical practice [6,10] and clinical epidemiology [2]. However, challenges related to data completeness and data quality need to be addressed to maximize the effectiveness of these efforts. For example bodyweight related measures (weight, height, abdominal circumference), are needed when clinicians calculate medication dosages based on body surface area (BSA) [11], or use body mass index (BMI) to estimate risk of cardiovascular disease, diabetes or cancer (Institute). Similarly, epidemiologists rely on bodyweight measures when determining novel associations such as the recently reported association between bodyweight and mortality due to influenza and pneumonia [4].

Despite the critical importance of bodyweight data for clinical care and research, several evaluations have pointed out that these

* Corresponding author at: Division of Epidemiology, University of Utah, 295 Chipeta Way, Salt Lake City, UT 84132, United States. Fax: +1 (801) 581 3623.
E-mail address: Maureen.Murtaugh@hsc.utah.edu (M.A. Murtaugh).

http://dx.doi.org/10.1016/j.jbi.2015.02.009
1532-0464/

data are frequently unavailable as structured data. For example researchers at the group health cooperative, testing the ability to use EHR data to calculate cardiac risk, found that among the records of 122,270 individuals, 11.5% were missing data for either height weight or both [5]. Similarly, Das et al. reported that among 1.8 Million Veterans who received outpatient care at VA facilities in the year 2000, 50.4% had no height or weight recorded as structured data [3]. More recently, Littman et al. reported that 32.8% of records of 173,127 veterans in the northwestern US were missing structured data for weight or height [7]. Since anecdotal reports suggested that in many cases individuals' heights and weights were measured during these visits, but the data was recorded as text in the clinical note, our research team felt that this was an important use case for information extraction.

In this project we sought to develop and validate a learning algorithm that would extract bodyweight related measures (weight, height, BMI, abdominal circumference) recorded in clinical notes from the VA's electronic Health record. We were motivated to explore this as an example of the potential to supplement structured data with data stored in text in order to fill in gaps in repeatedly measured clinical data. Our first aim was to determine how well we could capture weight, height, BMI and/or abdominal girth from outpatient notes. Our second aim was to determine the proportion of the data in the notes that was unique data.

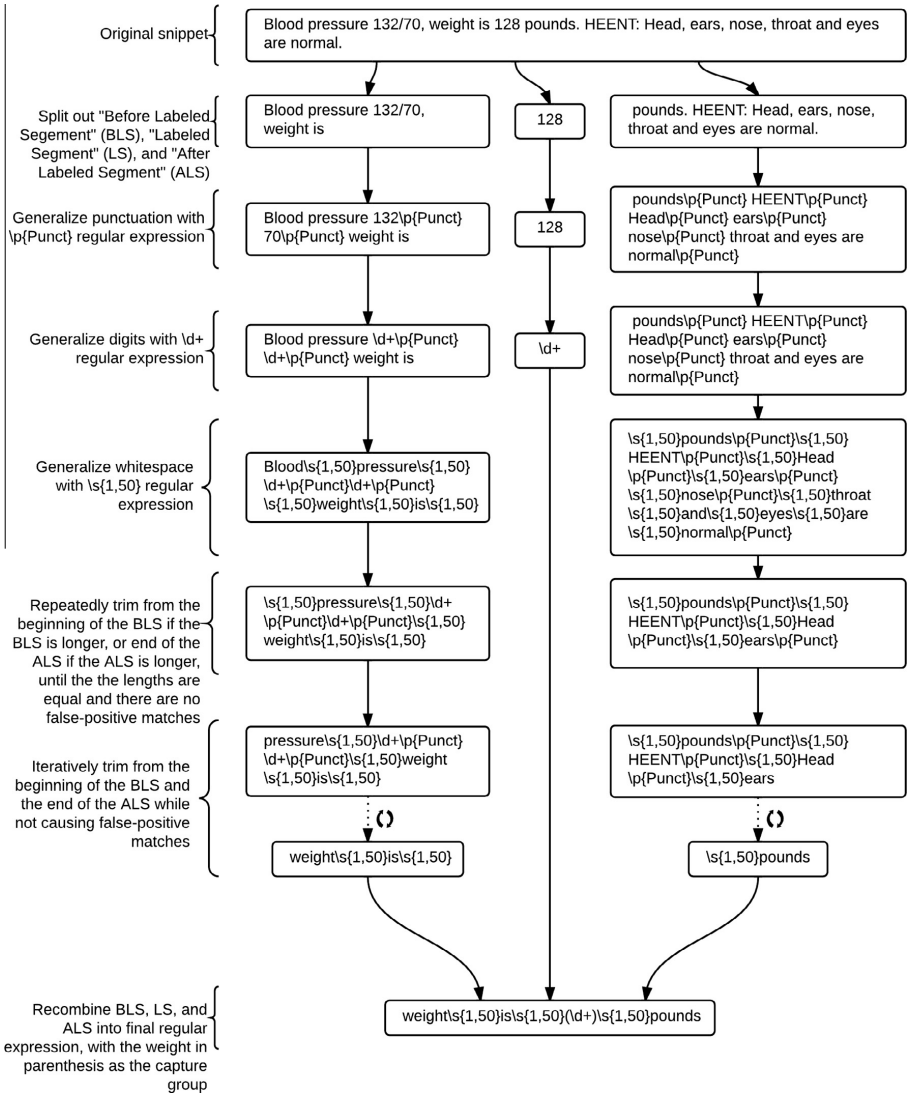

Fig. 1. Example of the creation of a standardized regular expression by REDEx.

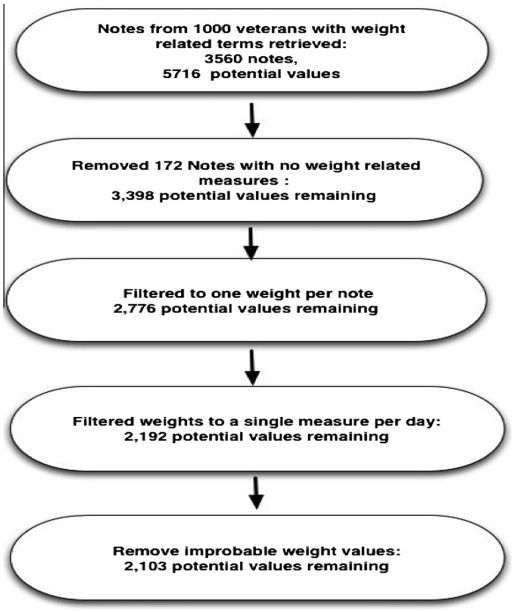

Fig. 2. Presents the data cleaning procedures used to ensure that we extracted only weight.

2. Background

2.1. Veterans Health Administration Informatics and Computing Infrastructure

The Veterans Health Administration was at the forefront of the development of Electronic Health Records and implemented its independently developed EHR, the Veterans Health Information Systems and Technology Architecture (Vista), in 1996. Therefore the Veterans Administration (VA) now has extensive longitudinal records on millions of Veterans.

Recognizing the opportunities for research using this aggregated data, the VA Health Services Research and Development (HSR&D) office funded the Veterans Informatics and Computing Infrastructure (VINCI) [12], a service-level collaboration between the Office of Information and Technology (OI&T) and the Office of Research and Development (OR&D). Designed to serve the data and Information technology needs of the VA research community, VINCI provides secure, centralized access to VA data resources in a high-performance computing environment. VINCI's mission is to provide researchers with an environment for efficient, secure analysis of patient level data, and to provide tools and coordination for research in basic and applied medical informatics.

As of FY 2013, VINCI provides access to structured and unstructured electronic medical data on 17,543,172 unique Veterans. The document corpus consists of 2,096,957,070 clinical documents from providers. The dataset also includes 1,611,284,360 diagnostic codes (ICD9), data on 1,654,598,048 pharmacy prescriptions, and 5,856,426,293 lab tests (both orders and results). Many other types of administrative and clinical data are also available.

2.2. Regular expression based learning

Regular expression-based learning has been an active area of research in computer science and to a lesser degree in biomedical informatics. Some learning algorithms require "seed" expressions, while others are designed to be totally automated. In the biomedical informatics domain, there is no completely automated learning algorithm for generating regular expressions that can be used to extract specific types of numerical values.

The goal of this project was to develop and test a Regular Expression Discovery extraction algorithm (REDEx, Fig. 1) that would address problems in typical regular expression based information extraction. First, the extraction of numerical values from clinical notes is typically performed using manually created regular expressions. This is a laborious process and its accuracy is dependent on the developers' expertise. Maintenance and extensions can also be particularly challenging as there is no standard method to document regular expressions and the patterns that match them. Finally, clinical domain experts who know best what they want to extract are generally not regular expressions experts, thus requiring additional labor to create libraries of regular expressions for more complex domains.

3. Materials and methods

We first retrieved and annotated a set of relevant outpatient notes as a reference standard. We then developed an NLP module for bodyweight related information extraction using the REDEx algorithm. Finally we applied the NLP module to a separate dataset to estimate the value of adding text data to structured data.

3.1. Retrieval of relevant notes

In order to collect the notes for annotation and use in the training of the classifier, we used Voogo [8,13], a search engine developed by our research group specifically to query VINCI data. It supports both free text and structured data searches and provides document, patient, and population-level results. Query results can be lists of patients and related documents, or summary reports that include the geographical distribution, age distribution, living/deceased status, gender, and prescribed VA medications for the veterans about whom the documents were written.

Table 1 provides the list of terms we used to retrieve notes that might contain bodyweight related measures. This list of terms was developed iteratively. We started with a preliminary set of search terms. Two annotators reviewed 268 notes that contained the initial terms. These terms were subsequently modified in order to ensure we were retrieving relevant notes (e.g. the initial search terms included abdominal, and abd* which resulted in many notes that referenced physical exam of the abdomen but were not relevant to our purpose). The annotators assessed the coverage of the initial terms in these notes. The final list of keywords was used to retrieve a second set of text snippets for annotation from 300

Table 1
Search terms used in Voogo to retrieve notes for training set.

Concept	Terms
Bodyweight	wt, weight, wgt, #, lb, kg
Height	height, ht, hgt
BMI	bmi, ibw, ibmi,
Abdominal Circumference	abdominal circumference, circumference, girth, waist circumference, whr, waist to hip ratio

notes and the test set of 3561 notes used to create the database (test set) of weights.

3.2. Annotation

We used an annotation tool developed at the National Libraries of Medicine (NLM) called VTT (Visual Tagging tool). This tool allows users to visually tag specific terms of interest in the clinical text that are instantiations of concepts of interest, the output of this tool includes unique identifiers for the notes containing the concept of interest, and the presence of specific tagged terms as coded data. Two researchers (BG and MM) independently annotated text snippets from 300 notes for the presence of the body weight-related measures of interest (weight, height, BMI, abdominal circumference). Text snippets are chunks of text of a limited length that may cross sentences, phrases or boundaries. The text snippets used in this study included the term of interest with a span of 20 words before and after.

Inter-annotator agreement between the two annotators varied based on the measure of interest but overall was excellent: the Kappa value for the weights extracted from the 968 snippets (extracted from 568 notes) was 99.54% for weight (kg, lbs, BMI, height, inches, cm, and feet) and 100% for abdominal circumference (included waist, cm inches, girth, $n = 22$). These annotated snippets were used to create the Regular Expression Discovery extraction algorithm (REDEx).

3.3. Regular Expression Discovery extraction algorithm (REDEx)

The REDEx algorithm builds upon our prior work [1] and contains the following main steps.

1. Each annotated snippet is split into parts: the labeled segment (LS, which is the text annotated by the researchers), before labeled segment (BLS, the text in the snippet preceding the LS), and after labeled segment (ALS, the text in the snippet following the LS).
2. The LS, BLS, and ALS are then converted into generalized regular expressions by first replacing all punctuation, digits, and whitespace with generalized expressions matching any punctuation, digits, or whitespace (e.g. "\p{Punct}", "\d+", or "\s{1,50}Br" respectively).
3. Each BLS-LS-ALS triplet is then progressively generalized by successively trimming from the front of the BLS and the end of the ALS until one or more false matches occurs.
4. Redundant triplets are then removed, and each remaining BLS-LS-ALS triplet is converted to a single regular expression, using the LS piece as the regular expression capture group.

The resultant set of regular expressions is then used as the "model" for subsequent extractions. Examples of snippets containing possible expression of weight are found in Appendix 2.

3.4. Final notes selection and data cleaning

We applied the REDEx algorithm to 3560 notes from 1000 Veterans selected at random spanning a period of October 1, 2011 through September 30, 2013 to identify 5716 probable weight values (Fig. 2). In this dataset of weight values extracted using REDEx, we identified 172 (of 5712) values that were obviously not weights. These appeared to be blood pressure (one value/another value), time (e.g. 1:00 PM), or a range of two values (e.g. 99–101, Appendix 1). We then filtered the values to include one per note. Next we filtered weights to include only one measurement per day, using the first measurement of the day when there were multiples (yielding 2192 values). Lastly, we cleaned values to remove weight values <75 and >600 lbs.

3.5. Confirmation of values from text and notes

In order to determine the concordance between bodyweight related values from text and the bodyweight values in the structured data, we calculated the correlation (Pearson's R) between pairs of values. The pairs included one that was taken from notes within one day of its' pair that was recorded as structured field data. We used SAS (Version 9.3) to calculate this correlation.

3.6. Estimation of the proportion of measure that were unique

To estimate the proportion of measures found by the REDex algorithm that were unique (not duplicating data stored in the structured data tables) we identified weights that were extracted from the text for which there was no related structured data within 1 day of the date of the note.

4. Results

4.1. Accuracy of extraction

The accuracy of REDEx was measured using annotations (actual values) from text snippets. A classification for a snippet was considered a true positive (TP) if the algorithm extracted a value that matched an actual value. A prediction was considered a false positive (FP) if the extractor yielded a value that did not match the actual value, or there was no actual value for the snippet. It was considered a false negative (FN) if the extractor did not predict a value when there was an actual value. It was considered a true negative (TN) when there was no predicted value and there was also no actual value for a snippet. Table 2 presents the confusion matrix for REDex evaluated using 968 snippets extracted from the 568 manually annotated notes. Evaluation was performed using 10-fold cross validation. Accuracy of the extractor was 98.3%, precision was 98.8%, recall was 98.3%, specificity was 98.1%, and *F*-score was 98.5%.

4.2. Confirmation of weights extracted from text

We used REDEx to select weight values from notes of Veterans who also had a weight in the structured field to confirm the reliability of the values extracted from text. The agreement of weights extracted from outpatient text notes with those extracted from the structured weight field within one day of each other was high (Pearson Correlation, $r = 0.95$).

4.3. Number of unique measures captured by algorithm

The proportion of weight values found by REDex that were unique measures was 162 of 2103 values, or 7.7%.

5. Discussion

In this paper we demonstrated that REDEx can be used to accurately find unique bodyweight related values in text. Compared to

Table 2
Confusion matrix for weight extractor applied to 968 snippets from 568 notes.

	Actual	
Predicted	584^{TP}	7^{FP}
	10^{FN}	367^{TN}

manually developing a set of regular expressions for the particular clinical task at hand, the automated learning algorithm provides a more generalizable solution. Extraction of these values had significant impact on both the numbers of unique individuals with bodyweight related measures and the numbers of measures for each individual. These finding suggest that this method could be used more generally for both clinical and research cohort identification to reduce missing data and to improve estimation of longitudinal trajectories of repeated measures within individuals.

Our findings echo other studies that have pointed to deficiencies in the availability of data in structured fields from electronic health records for bodyweight related measures. For example Green et al. reported that 11.5% of 122,270 patients were missing data in the EHR necessary to calculate BMI Similarly Rose et al. reported that among 79,947 patients served by a large primary care network, 39% (range 6–66%) did not have either height or weight recorded to allow for calculation of Body mass Index (BMI) [9]. The advantage of this study is that it takes the next step by developing an algorithm to extract data from clinical notes and address this missing data problem inherent in clinical care databases.

The impact of additional information from text notes is in part related to the magnitude of the Veterans Health Care system. For example, if this method were applied to the full VINCI database that represents 17 Million unique individuals and identified 7.7% new data points, it would identify 1,309,000 unique values. These values would provide data for individuals who would otherwise not have a weight related measure available for quantitative analysis within a 2-year time-period. Additionally, the methods would identify an average of two additional measures per person per year.

5.1. Strengths/limitations

This method could be applied to other variables in other corpus for a variety of clinical and research domains. For example, the method can be trained to recognize tumor margins or ejection fraction. Limitations include that the algorithm was trained to expressions appearing in the electronic health record at the VA. Importantly, REDEx allows flexibility in that it can be retrained if conventions for expressing the value of interest changes or are different in a different region or system. Methods that do not allow any false positives could lead to over fitting. In this particular use case though, we did not observe signs of over fitting in the expressions that were generated.

5.2. Future work

The extraction of these values enables us to evaluate a number of clinical interventions related to obesity and chronic disease. Future work can address the important question of whether the capture of this data would change the classification of individuals regarding their weight/obesity status and evaluate the impact on assessment of clinical outcomes related to changes in body weight or other vital signs. We are experimenting with permitting a small number of false positives in training, to avoid over-fitting and to tolerate a certain level annotation errors. Additionally, we will be able to demonstrate how the capture of data using NLP changes the estimation of trajectories in important vital sign data. Further, application of this method can be used to reduce potential sampling bias related to missing or mistimed clinical data.

6. Conclusion

A regular expression based learning algorithm to extract measures of interest in clinical text notes performed with high accuracy. The method extracts unique values (not stored else where on the clinical database) for Veterans who did have existing values. The method also identified additional values for Veterans who had some existing values. We believe this approach offers value for improving clinical data completeness and quality for both research and evaluating clinical care.

Acknowledgments

Funding for this project came from the following sources:

HIR 08-374: VA Health Services Research and Development. Consortium for Healthcare Informatics Research (CHIR).

HIR 08-204: Veterans Affairs Health Services Research & Development, VA Informatics and Computing Infrastructure (VINCI) aka Center for Scientific Computing.

CRE 12-315: Veterans Affairs Health Services Research & Development, CREATE: A VHA NLP Software Ecosystem for Collaborative Development and Integration.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2015.02.009.

References

[1] Bui DD, Zeng-Treitler Q. Learning regular expressions for clinical text classification. J Am Med Inform Assoc 2014;21:850–7.
[2] Chute CG. Invited commentary: observational research in the age of the electronic health record. Am J Epidemiol 2014.
[3] Das SR, Kinsinger LS, Yancy Jr W, Wang A, Ciesco E, Burdick M, et al. Obesity prevalence among veterans at Veterans Affairs medical facilities. Am J Prev Med 2005;28:291–4.
[4] Fisher-Hoch SP, Mathews CE, McCormick JB. Obesity, diabetes and pneumonia: the menacing interface of non-communicable and infectious diseases. Trop Med Int Heal 2013;18:1510–9.
[5] Green BB, Anderson ML, Cook AJ, Catz S, Fishman PA, McClure JB, et al. Using body mass index data in the electronic health record to calculate cardiovascular risk. Am J Prev Med 2012;42:342–7.
[6] Imler TD, Morea J, Imperiale TF. Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. Clin Gastroenterol Hepatol 2013.
[7] Littman A, Boyko EJ, McDonnell M, Fihn S. Evaluation of a weight managment program for veterans. Prev Chron Dis 2012;9:110267.
[8] Redd D, Rindflesch T, Nebeker J, Zeng-Treitler Q. Improve retrieval performance on clinical notes: a comparison of four methods. IEEE; 2013. p. 2389–97.
[9] Rose SA, Turchin A, Grant RW, Meigs JB. Documentation of body mass index and control of associated risk factors in a large primary care network. BMC Heal Serv Res 2009;9:236.
[10] Sohn S, Ye Z, Liu H, Chute CG, Kullo IJ. Identifying abdominal aortic aneurysm cases and controls using natural language processing of radiology reports. AMIA Summits Transl Sci Proc 2013;2013:249–53.
[11] Tipton P, Aigner M, Finto D, Haislet J, Pehl L, Sanford P, et al. Patient safety: consider the accuracy of height and weight measurements. Nursing 2012;42:50–2.
[12] US Department of Veterans Affairs. VA Informatics and Computing Infrastructure (VINCI) Washington DC: US Department of Veterans Affairs; 2013. <http://www.hsrd.research.va.gov/for_researchers/vinci/2013>.
[13] Zeng QT, Redd D, Rindflesch T, Nebeker, J. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. American Medical Informatics Association; 2012. 1050.